Hippocampus Segmentation using Patch-based Representation and ROC Label Enhancement

A.D. Tobar, J.C. Aguirre, D.A. Cárdenas-Peña, A.M. Álvarez-Meza, and C.G. Castellanos-Dominguez

Abstract—**Brain Magnetic Resonance Imaging (MRI) is a quantitative neuroimaging technique to support anatomical structure segmentation. Still, proper segmentation requires modeling small MRI regions, not to mention the class imbalance issue that yields false-positive predictions. This work introduces a Patch-based Segmentation with a Label Enhancement approach, termed PSLE, for improved MRI-based hippocampus segmentation, by combining different texture filters to capture salient patch relationships. First, we select target-related regions to perform a convex candidate combination for label estimation. Next, we improve the overall performance by fitting the decision threshold based on the Receiver Operating Characteristic (ROC) curve, tackling the class imbalance problem. In the middle, we analyze the effect on the performance metrics of the primary hyperparameters and stages (ablation study). Finally, the state-of-the-art methods are compared with multi-atlas segmentation and deep learning algorithms in three well-known hippocampus segmentation benchmark MRI collections: LONI, ADNI, and SATA.**

Index Terms—**Hippocampus segmentation, Texture features, Patch-based representation, Label enhancement, MRI.**

I. Introduction

Magnetic Resonance Imaging (MRI) segmentation is essential in clinical applications like detecting soft-tissue lesions and monitoring pathology progression [1]. Furthermore, for diseases like Alzheimer's, Schizophrenia, and Parkinson's, MRI segmentation assists the diagnosis and treatment by providing quantitative knowledge about the brain structures [2], [3], [4]. However, traditional manual delineation by expert clinicians is time and resource consumption, making such an approach unfeasible for large datasets and time-constrained applications [5]. Besides, the segmentation quality depends on the clinician's experience, knowledge, and carefulness [6], [7]. The above constraints and drawbacks made the development of automatic segmentation methodologies relevant for medical and image processing fields. Still, the fuzzy boundaries between structures challenge the automatic segmentation task due to their complex spatial dependencies, size and shape variance, and inhomogeneous magnetic properties [8].

Recent approaches address brain segmentation through deep learning (DL) [9]. Nevertheless, the memory requirement for learning 3D segmentation states an implementation drawback, even for models with a few layers. As an alternative, the DL approaches of Extended 2D Consensus Hippocampus Segmentation (E2D-Hipseg) [10], FastSurfer [11], Ataloglou [9], and Multi-Model Deep CNN (MMDC) [12] process the MRIs slice-wise along a given axis view. However, despite their suitable performance scores, such DL approaches demand a large amount of annotated data to yield reliable results in testing scenarios, which is hard to accomplish in medical imaging. Further, the small size of target structures yields class-imbalanced datasets that hamper the model's reliability [13], [14].

As a counterpart, Multi-Atlas Segmentation (MAS) lacks the above issues as it incorporates prior information through pre-segmented images, composing atlases from clinical knowledge as medical instances or templates [15]. MAS comprises two steps: firstly, image registration maps each atlas to the target image, and secondly, label fusion predicts each voxel's class. As a result, such approaches account for considerable anatomical variability from a few representative instances. Furthermore, the image processing in the target coordinate space raised an alternative MAS that weights templates according to their similarity to the target image, yielding subject-specific strategies that also tackle the intra-subject variability [16]. However, the dependence on the atlas-to-target registration reduces the segmentation accuracy in the presence of unrepresented anatomical differences that misaligns images [17].

Regarding this, patch-based segmentation strategies reduce the abovementioned issues by labeling voxels depending on intensity-based similarities over a predefined neighborhood. Some representative patch-based algorithms include the weighted voting strategy [18], the partially-localized random forests [19], the patch-wise metric learning with multi-scale features [20], and the labeling enhancement [21]. Nonetheless, as a price for the reduced registration dependence, most patch-based approaches demand more training voxels [22].

This work introduces a patch-based label fusion approach for MRI segmentation, considering enhanced specificity and sensitivity thresholding, termed Patch-based Segmentation with a Label Enhancement (PSLE). First, the proposal selects patches and extracts their multi-scale intensity and texture

Manuscript received July 21, 2022; revised Jaunary 6, 2023.

Under grants provided by the projects: "Herramienta de apoyo a la predicción de los efectos de anestésicos locales vía neuroaxial epidural a partir de termografía por infrarrojo" (Code 111984468021), funded by Minciencias; "Procesamiento de señales de eletroencefalografía en interfaz cerebro-computador orientado a la detección de imaginación motora utilizando modelos de aprendizaje profundo y medidas de conectividad" - (HERMES - 53223), funded by Universidad Nacional de Colombia; and UTP-6-21-5, funded by Universidad Tecnológica de Pereira.

A.D. Tobar is MSc student of the Universidad Nacional de Colombia, 170001 Manizales-Colombia (email: adtobarr@unal.edu.co)

J.C. Aguirre is MSc student of the Universidad Nacional de Colombia, 170001 Manizales-Colombia (email: jucaguirrear@unal.edu.co)

A.M. Álvarez-Meza is professor of Electrical, Electronics, and Computation Engineering Department, Universidad Nacional de Colombia, 170001 Manizales-Colombia (email: amalvarezme@unal.edu.co)

C.G. Castellanos-Dominguez is professor of Electrical, Electronics, and Computation Engineering Department, Universidad Nacional de Colombia, 170001 Manizales-Colombia (email: cgcastellanosd@unal.edu.co)

D.A Cárdenas-Peña is professor of Electrical Engineering Department, Universidad Tecnológica de Pereira, 60001 Pereira-Colombia (email: dcardenasp@utp.edu.co)

features from the original image space [8]. Afterward, a $k$nearest neighbor (kNN) classifier coarsely labels target voxels according to similarities in the feature space. In turn, the Received Operating Characteristic (ROC) curve allows for choosing the best-performed threshold to refine the coarse labeling [23]. Of note, hyperparameter tuning and ablation analysis are carried out. Mainly, we analyze the influence of the number of nearest neighbors for the kNN algorithm, the number of most similar atlases, and the patch and neighborhood radius. In addition, the ablation study evaluates the main model stages' influence on the final segmentation. Finally, performance comparison incorporates multi-atlas segmentation and deep learning algorithms in three well-known MRI-based hippocampus segmentation databases: LONI, ADNI, and SATA.

In a nutshell, our PSLE contribution is threefold: i) Employment of the Tanimoto similarity index on the original MRI intensity space to reject the unrepresentative patches [24]; ii) Augmented data representation using multi-scale and texture features to improve the discrimination between samples of different classes; and iii) Label enhancement based on ROC trade-off for MRI-based hippocampus segmentation.

The remainder of this paper is organized as follows: Section II formulates the proposed approach. Section III explores the experiments and results. Lastly, Section IV outlines the concluding remark.

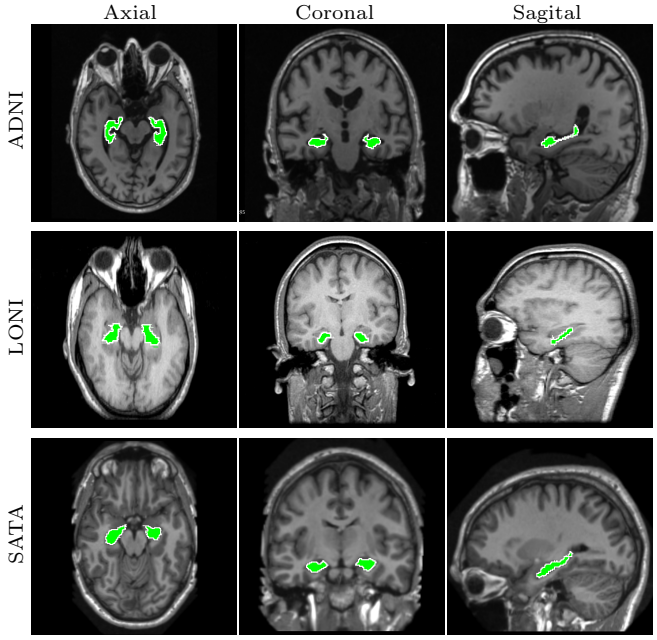


Fig. 1: Tested datasets for hippocampus segmentation. Some exemplary MRIs are depicted for ADNI, LONI, and SATA databases. Axial, Coronal, and Sagital views are presented for a given patient on each set, highlighting in green the region of interest concerning the hippocampus brain structure.

II. Materials and Methods

A. Datasets

This work considers the following datasets for assessing the hippocampus segmentation (see Figure 1):

- *Segmentation Algorithms, Theory and Applications (SATA)*: a publicly available collection created for assessing the image segmentation of blind-folded data. SATA holds 35 T1 MRIs-based manually delineated hippocampus.
- *Laboratory of Neuro-Imaging (LONI)* [25]: It assembles 40 T1-Weighted MRI brain images collected from healthy volunteers (20 males and 20 females), aging from 20 to 40 years.
- *Alzheimer's Disease Neuroimaging Initiative (ADNI)*: This database gathers imaging biomarkers, biosignals, and neuropsychological tests to characterize dementia patients. 100 T1 MRIs are selected from healthy subjects, holding their hippocampus masks. The subject's average age is 80 years old (49% male).

B. Patch-based Segmentation with Label Enhancement

Let $\mathcal{X} = \{x_t \in \mathbb{R} : t \in \Omega\}$ be an intensity MRI, being $x_t = \mathcal{X}(t)$ the voxel intensity at the t-th coordinate over the spatial domain Ω. Provided a set of tissue classes, the segmentation task assigns a label $l_t \in \mathcal{C}$ to the t-th voxel resulting in a semantic volume $\mathcal{L} = \{l_t \in \mathcal{C}\}$. Without loss of generality, this work states the binary case of hippocampus segmentation, so that $\mathcal{C} = \{0, 1\}$. The proposed methodology, termed Patch-based Segmentation with a Label Enhancement (PSLE), solves the hippocampus segmentation following the pipeline in Figure 2:

a) Preprocessing: that comprises motion correction and non-uniform intensity enhancement to mitigate blurring and Rician noise artifacts. The preprocessing stage maps MRIs to the Talairach space for spatial consistency over different atlases using histogram matching-based global and local affine registrations. Next, the Freesurfer and the BRAINSFit tools apply intensity normalization and skull stripping.

b) Atlas and voxel selection: uses the cross-modal mutual information (CMMI) measure to quantify the similarity between a target image $\mathcal{X}_*$ and the intensity-label atlas pair $\mathcal{A} = \{\mathcal{X}, \mathcal{L}\}$, as follows [26]:

$$\mathrm{CMMI}(\mathcal{X}_*, \mathcal{X}) = H(\mathcal{X}_*) + H(\mathcal{X}) - H(\mathcal{X}_*, \mathcal{X}), \quad (1)$$

where $H(\cdot)$ is the Shannon entropy operator. For coping with multi-atlas disagreement errors (different labels for a given voxel), the approach computes CMMI only at voxels with partial label agreement. Generally, such voxels belong to the tissue boundary and hold a high label and intensity variability. Label agreement is calculated as $\max_c \mathbb{E}_n\{P(l_t^n = c)\}$, where $P(l_t^n = c) \in [0, 1]$ is the class probability of atlas n at location t. $\mathbb{E}\{\cdot\}$ stands for the expected operator.

c) Patch-based feature extraction: gathers patch-label pairs $(\mathcal{P}_s, l_s)$ from T training atlases inside the neighborhood of the target voxel x_t, with $\mathcal{P}_s \in \mathbb{R}^{R\times R\times R}$ as the R-sided cubic patch centered at s. Further, a patch pre-selection procedure discards less reliable patches, with the advantage of decreasing the computational burden. The pre-selection employs the Tanimoto Similarity Index (TSI) due to its computational simplicity and benefits in quantifying data correspondence [24]:

$$\mathrm{TSI}(\mathbf{f}, \mathbf{f}') = \frac{\langle \mathbf{f}, \mathbf{f}' \rangle}{\langle \mathbf{f}, \mathbf{f} \rangle + \langle \mathbf{f}', \mathbf{f}' \rangle - \langle \mathbf{f}, \mathbf{f}' \rangle}, \quad (2)$$

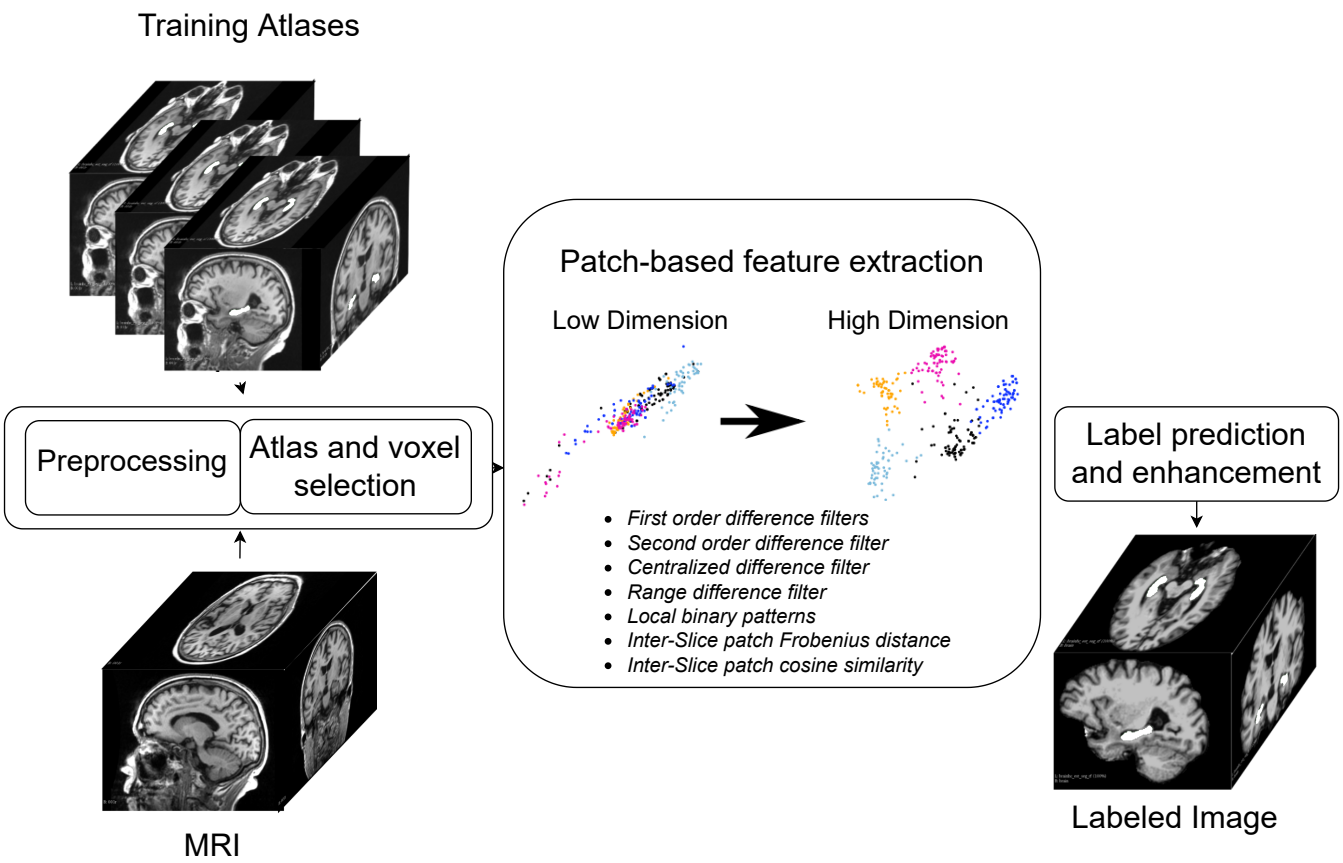


Fig. 2: PSLE pipeline for hippocampus segmentation.

being $\mathbf{f}, \mathbf{f}' \in \mathbb{R}^U$ a pair of feature vectors. Since patch intensity-based features lack the discrimination of adjacent structures sharing intensity patterns from MRIs [20], the pre-selection introduces the following texture and gradient features in multi-scaled intensity from a given image $\mathcal{X}$ [8].

First order difference filters (FODs) that are capable of detecting coarse intensity changes along a line segment without considering its central value:

$$\mathrm{FODs}(t, u) = \mathcal{X}(t + u) - \mathcal{X}(t - u), \tag{3}$$

where $u = (r\cos(\theta)\sin(\phi), r\sin(\theta)\cos(\phi), r\cos(\phi))$, $r \in \mathbb{R}^+$ is set as the patch radius, and the angles $\phi, \theta \in [-\pi, \pi]$ are fixed such that all elements in the patch are taken into account to compute the differences.

Second order difference filters (SODs) are employed to detect fine intensity changes along a line segment:

$$\mathrm{SODs}(t, u) = \mathcal{X}(t + u) + \mathcal{X}(t - u) - 2\mathcal{X}(t). \tag{4}$$

Centralized difference filters (CDFs) catch the signed distances from the central voxel and its surrounding neighbors:

$$\mathrm{CDFs}(t, t') = \mathcal{X}(t) - \mathcal{X}(t'),\ \forall t' \in \mathcal{Q}(t); \tag{5}$$

where $\mathcal{Q}(t)$ gathers the patch positions centered at t.

Centralized Euclidean distance (CED) captures the distance from the center to all voxels within $\mathcal{Q}(t)$:

$$\mathrm{CED}(t, t') = \sqrt{\sum_{\forall t' \in \mathcal{Q}(t)} (\mathcal{X}(t) - \mathcal{X}(t'))^2}. \tag{6}$$

Range difference filters (RDFs) compute the patch's range value centered at t:

$$\mathrm{RDFs}(t) = \max(\mathcal{P}(t)) - \min(\mathcal{P}(t)). \tag{7}$$

Local binary patterns (LBP) code center-referenced neighbor relations by generating a binary patch. Let $\mathscr{B}$ be a binary patch such that $\mathscr{B}(t') = 1$, if $\mathcal{P}(t') \geq \mathcal{P}(t)$; otherwise, $\mathscr{B}(t') = 0$. Then, LBP can be obtained by the following linear combination:

$$\mathrm{LBP}(t) = \mathrm{vec}(\mathscr{B})^\top \mathrm{vec}\left(\{2^i\}_{i=0}^{|\mathscr{B}|}\right). \tag{8}$$

Inter-Slice patch Frobenius distance (ISFD) obtains the distance from two Saggital slices aside t position:

$$\mathrm{ISFD}(t, \tilde{u}) = \|S(t - \tilde{u}) - S(z + \tilde{u})\|_F, \tag{9}$$

being $\tilde{u} = (0, 0, 1)$, $S(t)$ is a 2D Saggital patch slice centered at t, and $||\cdot||_F$ is the Frobenius norm.

Inter-Slice patch cosine similarity (ISCS) calculates a normalized positive defined similarity from two Saggital slices aside t position, yielding:

$$\mathrm{ISCS}(t, \tilde{u}) = \frac{\sqrt{\mathrm{tr}(S^\top(t - \tilde{u})S(t + \tilde{u}))}}{\|S(t - \tilde{u})\|_F \|S(t + \tilde{u})\|_F}. \tag{10}$$

The outputs of all the above descriptors are concatenated to form a feature vector $\mathbf{f}_t$.

d) Label prediction and ROC-based enhancement: This stage segments a query MRI by solving a voxel-wise classification problem using a k-nearest neighbors (k-NN)

classifier. It fuses candidate labels into likelihoods $\hat{p}_t \in \mathbb{R}^+$ using the normalized convex label combination:

$$\hat{p}_t = \frac{\sum_{k=1}^{K} w_{t,k} y_k}{\sum_{i=1}^{K} w_{t,k}}, \tag{11}$$

where $y_k \in \mathcal{C}$ holds the k-th neighbor label and weights results from the RBF kernel as $w_{t,k} = \exp\left(\frac{-\|\mathbf{f}_t - \mathbf{f}_k\|_2^2}{2\sigma^2}\right)$, with $\sigma \in \mathbb{R}^+$ as the kernel bandwidth. Lastly, an ϵ-based thresholding yields the hard label prediction:

$$\hat{l}_t = \begin{cases} 1 & \text{if } \hat{p}_t \geq \epsilon \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

The $\epsilon \in \mathbb{R}^+$ threshold plays an essential role in label prediction. Its typical value of 0.5 assumes balanced classes, which underperforms for complex shapes and relatively small structures, i.e., the hippocampus. Therefore, this work proposes to maximize the True Positive Rate (TPR) over the False Positive Rate (FPR) ratio w.r.t. the threshold as:

$$\epsilon^\star = \arg\max_{\epsilon} \frac{TPR(\epsilon)}{FPR(\epsilon)}. \tag{13}$$

e) Tuning and validation: A leave-one-subject-out strategy assesses the segmentation performance in terms of four different metrics: *Dice Similarity Index* (DSI), *Jaccard Similarity Index* (JSI), *Precision Index* (PSI), and *Recall Index* (REC) [23].

III. Results and Discussion

This section presents the methodology results in terms of interpretability and segmentation performance. For comparison purposes, this work considers the four MAS techniques *Majority Voting* (MV) [27], *Local Weighted Voting* (LWV) [28], *Non-Local Weighted Voting* (NLWV) [29], and *Joint Label Fusion* (JOINT) [30], and two DL models E2DHipseg [10] and FastSurfer [11] . PSLE, MV, LWV, and NLWV were implemented using the Insight Toolkit (ITK-5.0) and the linear algebra library Eigen-3.3.9, while JOINT using the Advanced Normalization Tools package. Namely, several performance metrics describe the influence of the hyperparameter tunning over the PLSE, and an interpretability analysis quantifies the effect of each proposed stage on each dataset. Finally, we compare state-of-the-art hippocampus segmentation models against our proposed PLSE.

A. PSLE Hyperparameter Analysis

Figure 3 presents the PSLE segmentation performance concerning the number of nearest neighbors for the k-NN classifier. At first glance, the larger the number of neighbors, the larger the dispersion of the performance metric. Further, the JSI evidences that the false positive rate grows faster than the false negative rate with the number of neighbors. Hence, including fewer related candidate labels decreases the segmentation quality. We fix the neighbors to 12, 17, and 7 for SATA, LONI, and ADNI, respectively.

Likewise, Figure 4 illustrates the influence of the number of atlases, path and neighborhood radius size, and the ROC thresholding for segmentation performance enhancement for all considered datasets. Figure 4a presents the influence of the number of selected atlases in PSLE, ranging from 2 to 25.

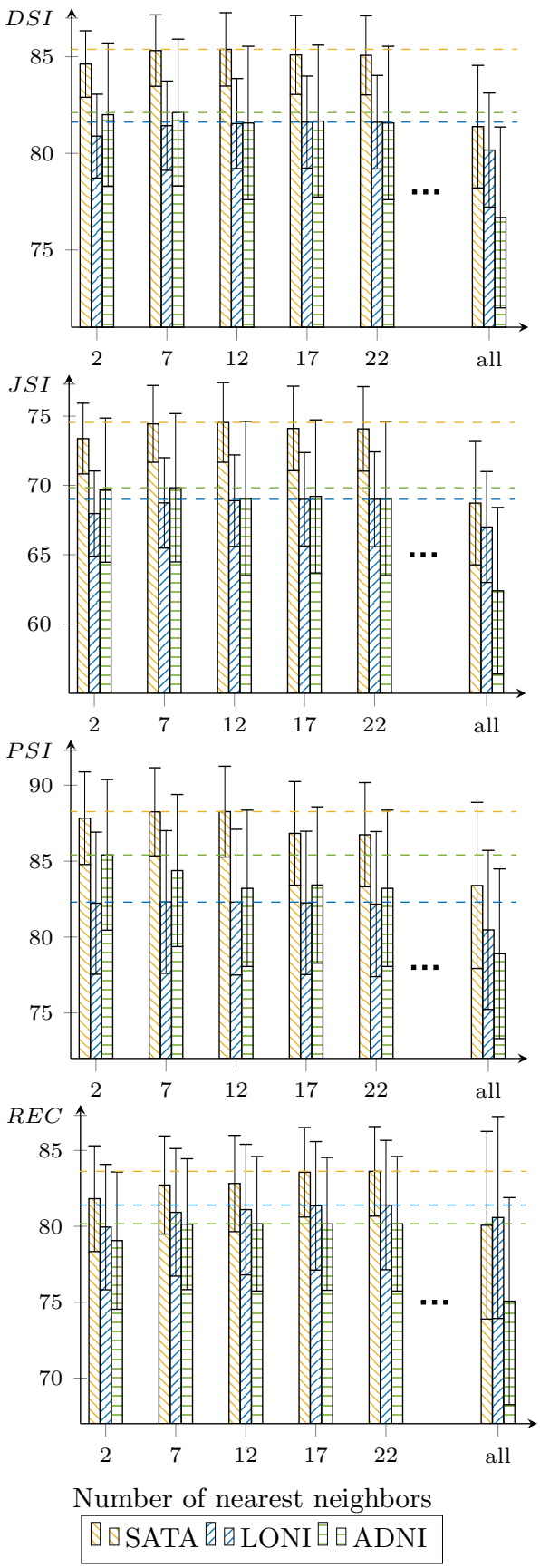


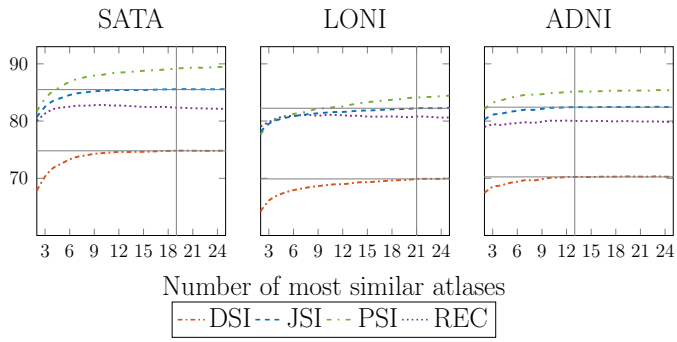
Fig. 3: Number of nearest neighbors for the kNN classifier. We searched $k \in \{2, 7, 12, 17, 22, all\}$, where all refers to the total number of training samples for the given test voxel, showing REC, PSI, JSI, and DSI indices for hippocampus segmentation in three different datasets, SATA, LONI and ADNI, plotted versus the number of nearest neighbors.

Figure 3a highlights that in most cases, PSLE early reaches the maximum values for both JSI and DSI indices. This fact suggests that too many atlases, instead of producing a performance enhancement, detriment the labeling quality due to unrepresentative atlases introducing noise. On the other hand, REC and PSI curves describe an opposed behavior as the number of atlases rises; that is, PSI increases, whereas REC drops. These tendencies indicate a shrunk segmentation volume for larger atlas sets, also a hint of under-segmentation in the presence of anatomical variability.

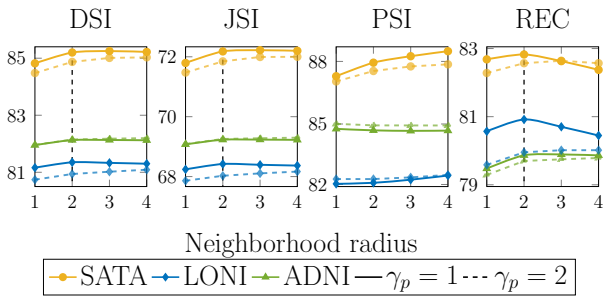
Regarding the patch and neighborhood radius tuning, the optimal radii depend on the labeling method, atlas selection, and registration. We assess the PSLE performance for a patch radius size of $\gamma_p \in \{1, 2\}$ (ranging from $3 \times 3 \times 3$ till $5 \times 5 \times 5$ voxels) and neighborhood radius size $\gamma_n \in \{1, 2, 3, 4\}$ (ranging from $3 \times 3 \times 3$ till $9 \times 9 \times 9$ voxels). Figure 4b depicts DSI and JSI reaching the maximum values at a neighborhood radius equal to 2 (a total volume of $5 \times 5 \times 5$). Further, they remain relatively constant, meaning that the misalignment produced by the affine registration demands a larger explored neighborhood. Contrarily, large patches miss fine tissue properties.

Since the proposed methodology aims at the ROC-based label enhancement, PSLE is tested by varying the ϵ threshold within the range $[0.1, 0.9]$ to find a trade-off between model sensitivity and specificity. As evidenced in Figure 4c, PSLE attains suitable ROC curves for all considered databases, with Area Under the ROC (AUC) values larger than 80%. Further, the optimal segmentation threshold, denoted as a bullet mark, results in $\epsilon^\star$ as 0.41, 0.35, and 0.43 for SATA, LONI, and ADNI, respectively. Note that optimal thresholds are far from the standard value of 0.5 for label estimation, proving the benefits of the proposed ROC-based label enhancement.

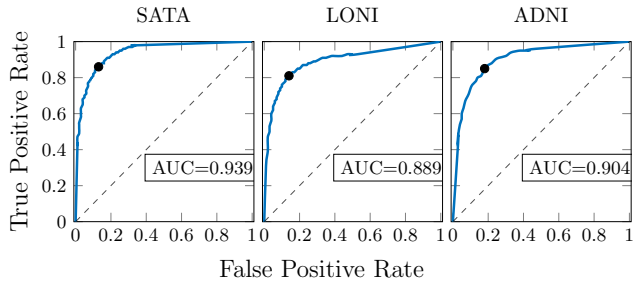
For visual assessment, Figure 5presents the best and worst segmentations for PLSE, E2Dhipseg, and FastSurfer. In general, mislabeling arise around the tissue boundaries as initially expected. Nonetheless, PLSE stands out as the most consistent approach through all datasets, as it performs comparably better in its worst cases than state-of-the-art approaches, evidencing a generalization gain. Further, the worst-segmented subjects and the best in LONI exhibit an under-segmentation that can be effortlessly explained owing to the sensitive nature of the manually labeled region. Contrarily, the best-labeled SATA and ADNI subjects depict minor errors, equally distributed in false positives and false negatives. Therefore, the proposed PSLE is more reliable as errors evenly distribute over the datasets, images, and performance metrics.



(a) Number of considered atlases.



(b) Patch and neighborhood radius size.



(c) Thresholding and ROC-based enhancement.

Fig. 4: Performance metrics for the PSLE hyperparameter analysis as a function of the number of considered atlases (top), patch and neighborhood radius (middle), and labeling threshold (bottom).

B. Ablation Study

We analyze the influence of the PSLE stages (see Figure 2) on hippocampus segmentation performance for fixed input parameters. Table I quantitatively evaluates the effect of progressively introducing each stage into the processing pipeline. The most evident result is the successive improvement in performance for SATA and ADNI while including each stage, resulting in an average enhancement of around 3% and 5%, respectively. Instead, LONI describes an odd behavior due to the preselection step that reduces the improvement of the kNN-based estimator, suggesting a lack of heterogeneity in training samples for most target voxels. For example, the feature extraction stage decreases recall by 12% but increases precision by about 10%. It means that segmented volume size decrement tremendously by adding texture and gradient features, which could be a consequence of the manual segmentation sharpness in the database. Finally, the joining of ROC-based label correction discovers a proper trade-off between REC and PSI indices. Such compensation allows an enhancement of around half percent in DSI and JSI indices while a 4% in the recall. In general, each of the introduced stages provides a complementary alternative for improving segmentation performance in both average and standard deviation for all metrics and databases, becoming a progressively strengthened segmentation approach.

C. Method Comparison Results

Finally, Table II compares the segmentation metrics of considered approaches. Overall metrics, the proposed PSLE outperforms other atlas-based strategies, thanks to the ROC-based label enhancement stage. Particularly for precision, JOINT improves PSLE at the cost of an over-segmentation.

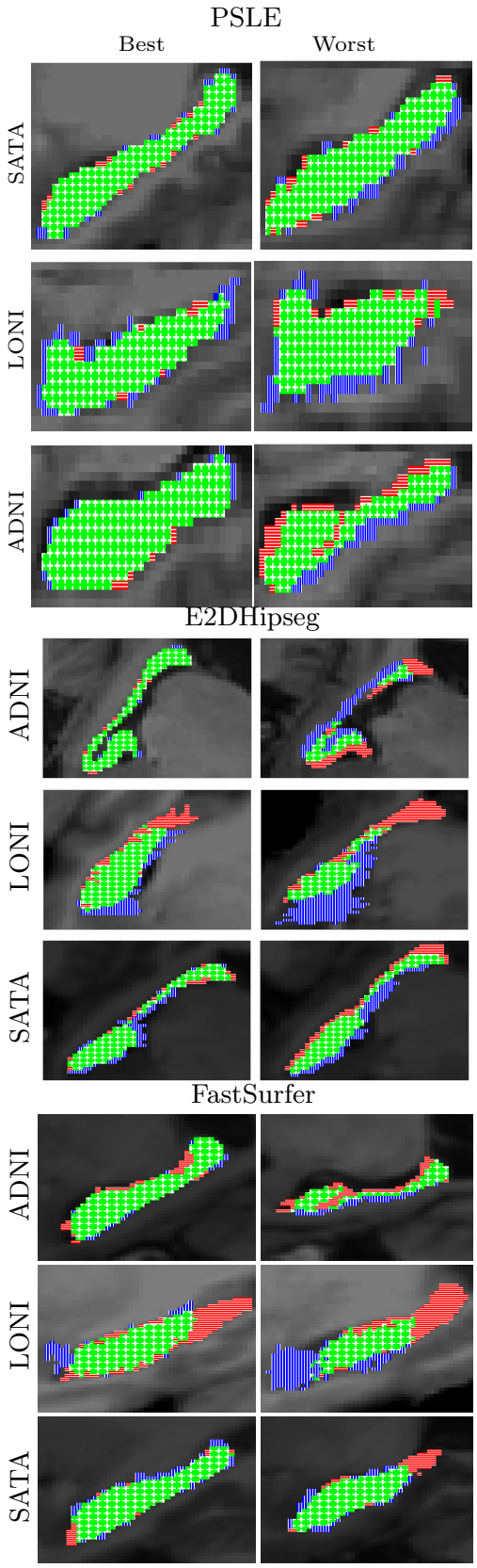


Fig. 5: Visual inspection results. PSLE (ours), E2DHipseg, and FastSurfer approaches are compared. The best (top row) and worst (bottom row) samples are presented for each database (SATA, LONI, and ADNI), denoting in green, blue, and red the correctly labeled (true positives), under-segmented (false negatives), and over-segmented (false positives) voxels, respectively.

TABLE I: Influence of PSLE stages in the hippocampus segmentation performance. Pre-selection, Patch-based feature extraction, and ROC-based estimation correction are studied.

Method	DSC	JSI	PSI	REC
SATA				
Baseline	84.0 ± 2.56	72.5 ± 3.79	83.5 ± 4.07	84.6 ± 3.60
Preselection	84.5 ± 2.27	73.3 ± 3.41	85.1 ± 3.35	84.2 ± 3.64
Feature extraction	85.0 ± 2.05	74.0 ± 3.10	86.1 ± 3.06	84.2 ± 3.31
ROC enhancement	86.3 ± 1.83	76.0 ± 2.83	85.3 ± 2.58	87.4 ± 2.53
LONI				
Baseline	80.7 ± 2.43	67.8 ± 3.41	79.2 ± 5.52	82.9 ± 4.23
Preselection	80.7 ± 2.43	67.7 ± 3.39	79.2 ± 5.57	82.7 ± 4.41
Feature extraction	78.4 ± 2.45	64.5 ± 3.28	89.0 ± 4.21	70.3 ± 4.87
ROC enhancement	83.0 ± 2.12	70.9 ± 3.07	80.7 ± 3.47	85.4 ± 2.30
ADNI				
Baseline	77.7 ± 5.66	63.9 ± 7.37	76.8 ± 6.77	78.8 ± 5.81
Preselection	78.0 ± 5.44	64.2 ± 7.10	77.1 ± 6.47	79.1 ± 5.78
Feature extraction	80.9 ± 4.90	68.2 ± 6.63	80.1 ± 6.03	81.9 ± 4.69
ROC enhancement	82.6 ± 4.57	70.6 ± 6.33	81.2 ± 6.13	84.1 ± 3.88

Regarding the DL approaches, E2DHipseg reaches the highest scores in the ADNI dataset, mainly because the deep network was trained on such a large dataset. Nonetheless, E2DHipposeg underperforms the MAS approaches in SATA and LONI, proving the lack of generality. Similarly, FastSurfer reaches its best performance on the SATA dataset, but loses considerable performance over LONI dataset. Hence, the proposed PSLE yields the most balanced performance with an improved generalization of the segmentation task.

TABLE II: MRI-based hippocampus segmentation results. Method comparison is shown concerning the mean ± standard deviation of the considered performance measures.

Method	DSI	JSI	PSI	REC
SATA				
MV [27]	77.3 ± 4.72	63.2 ± 6.10	81.2 ± 6.40	73.9 ± 4.92
LWV [28]	78.7 ± 5.04	65.2 ± 6.63	81.0 ± 6.61	76.8 ± 5.20
NLWV [29]	78.8 ± 4.72	65.3 ± 6.24	83.6 ± 6.18	74.9 ± 5.73
JOINT [30]	85.1 ± 2.54	74.2 ± 3.81	88.9 ± 3.13	81.8 ± 3.57
E2DHipseg[10]	78.6 ± 0.26	64.8 ± 0.35	84.8 ± 0.37	73.3 ± 0.28
FastSurfer [11]	80.01 ± 0.17	66.8 ± 0.24	89.9 ± 0.31	72.3 ± 0.25
PSLE (ours)	86.3 ± 1.83	76.0 ± 2.83	85.3 ± 2.58	87.4 ± 2.53
LONI				
MV [27]	79.6 ± 2.40	66.2 ± 3.30	82.7 ± 4.80	77.1 ± 4.20
LWV [28]	80.1 ± 2.23	66.8 ± 3.22	82.5 ± 4.65	78.1 ± 4.13
NLWV [29]	79.9 ± 2.42	66.6 ± 3.34	83.4 ± 4.46	77.1 ± 4.86
JOINT [30]	82.9 ± 2.10	70.9 ± 3.06	86.2 ± 4.28	80.1 ± 3.64
E2DHipseg[10]	52.6 ± 0.41	35.8 ± 0.38	66.5 ± 0.44	43.1 ± 0.51
FastSurfer [11]	57.0 ± 0.29	39.9 ± 0.28	58.7 ± 0.31	55.5 ± 0.45
PSLE (ours)	83.0 ± 2.12	70.9 ± 3.07	80.7 ± 3.47	85.4 ± 2.30
ADNI				
MV [27]	74.9 ± 5.44	60.2 ± 6.88	78.1 ± 7.49	72.4 ± 6.23
LWV [28]	75.8 ± 5.28	61.3 ± 6.78	78.5 ± 7.46	73.6 ± 5.67
NLWV [29]	74.3 ± 4.98	59.4 ± 6.25	80.50± 6.52	69.5 ± 6.02
JOINT [30]	81.0 ± 5.55	68.5 ± 7.42	84.7 ± 6.44	77.8 ± 6.05
E2DHipseg[10]	90.0 ± 1.00	84.5 ± 1.00	90.7 ± 0.7	91.8 ± 0.85
FastSurfer [11]	79.0 ± 0.42	65.5 ± 0.50	71.2 ± 0.50	89.0 ± 0.30
PSLE (ours)	82.5 ± 4.57	70.5 ± 6.33	81.2 ± 6.13	84.1 ± 3.88

IV. Conclusions

We introduce an MRI-based hippocampus segmentation approach named Patch-based Segmentation with a Label Enhancement (PSLE). Our proposal employs a multi-atlas strategy coupled with multi-scale and texture features to reveal relevant relationships among subjects and patches. Besides, PSLE allows coding essential inner-patch information that enables proper candidate label combination and ROC-based enhancement to deal with unbalanced classes,

outliers, and edge sharpness. Moreover, the experimental analysis of the primary PSLE hyperparameters yields a suitable performance for three well-known MRI databases, e.g., SATA, LONI, and ADNI. The further ablation study demonstrated that each considered stage complementary supports the hippocampus segmentation. In particular, ROC-based label enhancement discovers a proper trade-off between REC and PSI indices, supporting hippocampus detection. In turn, method comparison included multi-atlas and deep learning algorithms. Obtained results proved PSLE as a suitable multi-atlas alternative for brain structure segmentation from MRIs, outperforming multi-atlas approaches and being competitive against more elaborated deep learning methods.

For future work, the authors plan to join the ROC-based enhancement within DL strategies to take advantage of the data-driven feature extraction without hampering generalization. Then, we will couple loss functions devoted to imbalance classification [31] and DL approaches for feature representation [32] to improve boundary delineation of small structures.

REFERENCES

[1] M. Z. Khan, M. K. Gajendran, Y. Lee, and M. A. Khan, "Deep neural architectures for medical image semantic segmentation: Review," *IEEE Access*, vol. 9, no. 1, pp. 83 002–83 024, 2021.

[2] J. Shi, R. Zhang, L. Guo, L. Gao, H. Ma, and J. Wang, "Discriminative feature network based on a hierarchical attention mechanism for semantic hippocampus segmentation," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 2, pp. 504–513, 2021.

[3] A. Desiani, B. Suprihatin, S. Yahdin, A. I. Putri, F. R. Husein *et al.*, "Bi-path architecture of cnn segmentation and classification method for cervical cancer disorders based on pap-smear images." *IAENG International Journal of Computer Science*, vol. 48, no. 3, pp. 782–791, 2021.

[4] M. Marwan, F. AlShahwan, F. Sifou, A. Kartit, and H. Ouahmane, "Improving the security of cloud-based medical image storage." *Engineering Letters*, vol. 27, no. 1, pp. 175–193, 2019.

[5] R. Yang and Y. Yu, "Artificial convolutional neural network in object detection and semantic segmentation for medical imaging analysis," *Frontiers in Oncology*, vol. 11, no. 1, p. 573, 2021.

[6] Z. Z. Zhijun Luo and B. Zhang, "An scadtv nonconvex regularization approach for magnetic resonance imaging," *IAENG International Journal of Computer Science*, vol. 48, no. 4, pp. 1013–1020, 2021.

[7] D. Karimi, H. Dou, S. K. Warfield, and A. Gholipour, "Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis," *Med. Image Anal.*, vol. 65, no. 1, p. 101759, 2020.

[8] P. Yi, L. Jin, T. Xu, L. Wei, and G. Rui, "Hippocampal segmentation in brain mri images using machine learning methods: A survey," *Chinese Journal of Electronics*, vol. 30, no. 5, pp. 793–814, 2021.

[9] D. Ataloglou, A. Dimou, D. Zarpalas, and P. Daras, "Fast and precise hippocampus segmentation through deep convolutional neural network ensembles and transfer learning," *Neuroinformatics*, vol. 17, no. 4, pp. 563–582, 2019.

[10] D. Carmo, B. Silva, C. Yasuda, L. Rittner, and R. Lotufo, "Hippocampus segmentation on epilepsy and alzheimer's disease studies with multiple convolutional neural networks," *Heliyon*, vol. 7, no. 2, p. e06226, 2021.

[11] L. Henschel, S. Conjeti, S. Estrada, K. Diers, B. Fischl, and M. Reuter, "Fastsurfer - a fast and accurate deep learning based neuroimaging pipeline," *NeuroImage*, vol. 219, no. 2, p. 117012, 2020.

[12] M. Liu, F. Li, H. Yan, K. Wang, Y. Ma, L. Shen, and M. Xu, "A multi-model deep convolutional neural network for automatic hippocampus segmentation and classification in alzheimer's disease," *NeuroImage*, vol. 208, no. 1, p. 116459, 2020.

[13] E. Tappeiner, M. Welk, and R. Schubert, "Tackling the class imbalance problem of deep learning-based head and neck organ segmentation," *International Journal of Computer Assisted Radiology and Surgery*, vol. 17, no. 11, pp. 2103–2111, 2022.

[14] S. Asgari, K. Abhishek, J. Cohen, J. Cohen-Adad, and G. Hamarneh, "Deep semantic segmentation of natural and medical images: a review," *Artificial Intelligence Review*, vol. 54, no. 1, pp. 137–178, 2021.

[15] Y. Zhang, J. Duan, Y. Sa, and Y. Guo, "Multi-atlas based adaptive active contour model with application to organs at risk segmentation in brain mr images," *IRBM*, vol. 43, no. 3, pp. 161–168, 2022.

[16] L. Sun, W. Shao, M. Wang, D. Zhang, and M. Liu, "High-order feature learning for multi-atlas based label fusion: Application to brain segmentation with mri," *IEEE Transactions on Image Processing*, vol. 29, no. 1, pp. 2702–2713, 2020.

[17] M. Antonelli, M. J. Cardoso, E. W. Johnston, M. B. Appayya, B. Presles, M. Modat, S. Punwani, and S. Ourselin, "Gas: A genetic atlas selection strategy in multi-atlas segmentation framework," *Medical Image Analysis*, vol. 52, no. 1, 2019.

[18] J. E. Iglesias and M. R. Sabuncu, "Multi-atlas segmentation of biomedical images: a survey," *Medical image analysis*, vol. 24, no. 1, pp. 205–219, 2015.

[19] Q. Zheng and Y. Fan, "Integrating semi-supervised label propagation and random forests for multi-atlas based hippocampus segmentation," in *2018 IEEE 15th International Symposium on Biomedical Imaging, ISBI 2018*, Washington, DC, USA, 4-7 April, 2018, pp. 154–157.

[20] Y. Wang, G. Ma, X. Wu, and J. Zhou, "Patch-based label fusion with structured discriminant embedding for hippocampus segmentation," *Neuroinformatics*, vol. 16, no. 3-4, pp. 411–423, 2018.

[21] L. Sun, C. Zu, W. Shao, J. Guang, D. Zhang, and M. Liu, "Reliability-based robust multi-atlas label fusion for brain mri segmentation," *Artificial intelligence in medicine*, vol. 96, pp. 12–24, 2019.

[22] D. Cárdenas-Peña, A. Tobar-Rodríguez, G. Castellanos-Domínguez, and A. D. N. Initiative, "Adaptive Bayesian label fusion using kernel-based similarity metrics in hippocampus segmentation," *Journal of Medical Imaging*, vol. 6, no. 1, p. 014003, 2019.

[23] A. Geron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*, 2nd ed. O'Reilly Media, Inc., 2019.

[24] D. C. Anastasiu and G. Karypis, "Efficient identification of tanimoto nearest neighbors," *International Journal of Data Science and Analytics*, vol. 4, no. 3, pp. 153–172, 2017.

[25] D. W. Shattuck, M. Mirza, V. Adisetiyo, C. Hojatkashani, G. Salamon, K. L. Narr, R. A. Poldrack, R. M. Bilder, and A. W. Toga, "Construction of a 3d probabilistic atlas of human cortical structures," *Neuroimage*, vol. 39, no. 3, pp. 1064–1080, 2008.

[26] X. Zhuang and J. Shen, "Multi-scale patch and multi-modality atlases for whole heart segmentation of mri," *Medical image analysis*, vol. 31, no. 1, pp. 77–87, 2016.

[27] R. A. Heckemann, J. V. Hajnal, P. Aljabar, D. Rueckert, and A. Hammers, "Automatic anatomical brain mri segmentation combining label propagation and decision fusion," *NeuroImage*, vol. 33, no. 1, pp. 115–126, 2006.

[28] X. Artaechevarria, A. Munoz-Barrutia, and C. Ortiz-de Solórzano, "Combination strategies in multi-atlas image segmentation: application to brain mr data," *IEEE transactions on medical imaging*, vol. 28, no. 8, pp. 1266–1277, 2009.

[29] P. Coupé, J. V. Manjón, V. Fonov, J. Pruessner, M. Robles, and D. L. Collins, "Patch-based segmentation using expert priors: Application to hippocampus and ventricle segmentation," *NeuroImage*, vol. 54, no. 2, pp. 940–954, 2011.

[30] H. Wang, J. W. Suh, S. R. Das, J. B. Pluta, C. Craige, and P. A. Yushkevich, "Multi-atlas segmentation with joint label fusion," *IEEE PAMI*, vol. 35, no. 3, pp. 611–623, 2013.

[31] C. Jimenez-Castano, A. Alvarez-Meza, and A. Orozco-Gutierrez, "Enhanced automatic twin support vector machine for imbalanced data classification," *Pattern Recognition*, vol. 107, no. 1, p. 107442, 2020.

[32] C. Jimenez-Castaño, A. Álvarez Meza, O. Aguirre-Ospina, D. Cárdenas-Peña, and A. Orozco-Gutierrez, "Random fourier features-based deep learning improvement with class activation interpretability for nerve structure segmentation," *Sensors*, vol. 21, no. 22, p. 7741, 2021.