Journal of Science and Technology – Smart Systems and Devices
Volume x, Issue x, Month Year, ppp-ppp

NeuroAnatomy-Aware Refinement in Intracerebral Hemorrhage Segmentation and Towards Traumatic Brain Injury Severity Assessment

Hoang Bach Nguyen[1], Quang Tung Pham[1], Sinh Huy Nguyen[1], Chi Thanh Nguyen[1], Thanh Hai Tran[2], Van Tuan Nguyen[3], Hai Vu[2,*]

[1]*Academy of Military Science and Technology, Hanoi, Vietnam*
[2]*Hanoi University of Science and Technology, Hanoi, Vietnam*
[3]*Institute of Diagnostic and Interventional Radiology, Bach Mai Hospital, Hanoi, Vietnam*
[*]*Corresponding author email: hai.vu@hust.edu.vn*

Abstract

Timely and objective traumatic brain injury (TBI) severity assessment is essential for clinical decision-making and for integrating imaging into digital health workflows. This paper presents a multimodal framework that combines quantitative biomarkers derived from non-contrast head CT with structured clinical variables to predict TBI severity. On the imaging branch, a slice-based segmentation approach is adopted to handle heterogeneous CT volumes with varying numbers of axial slices, together with an anatomy-aware refinement step to improve label consistency and anatomical plausibility of hemorrhage recognition. The extracted imaging biomarkers are then integrated with key clinical indicators and used for severity stratification by conventional machine-learning classifiers. Experiments on a matched multimodal dataset consisting of head CT images and structured clinical and tabular data demonstrate that incorporating segmentation-derived imaging features improves the prognostic assessment of TBI severity compared with using clinical variables alone. These findings highlight the added value of lesion quantification from CT for multimodal severity prediction and support the practicality of the proposed framework for clinical severity assessment.

Keywords: CT segmentation, digital health, imaging biomarkers, traumatic brain injury.

1. Introduction

Traumatic brain injury (TBI) is a major global health burden affecting about 369 per 100,000 individuals and causing 30–50% of trauma-related deaths [1]. TBI severity is commonly categorized into four levels (mild, moderate, severe, critical) using clinical indicators such as the Glasgow Coma Scale (GCS), which guide treatment and prognosis. Early and accurate severity assessment is vital for appropriate interventions, especially in severe cases where timely decisions affect survival [2]. Despite advances in machine learning (ML) for TBI prognosis [2], challenges remain due to heterogeneous datasets, non-standardized benchmarks, class imbalance, and limited use of imaging biomarkers [3]. Current workflows still depend on bedside scales and qualitative CT readings, which may overlook subtle pathology. ML models further struggle with irrelevant features, imbalance, and missing data, while most focus on outcome prediction rather than interpretability.

Another practical challenge in traumatic brain injury research is the limited availability of matched multimodal datasets that simultaneously provide head CT images, structured clinical variables, and sufficiently detailed imaging annotations for integrated severity assessment. Existing public resources often cover only part of this requirement. For example, the RSNA 2019 Brain CT Hemorrhage dataset [4] is a large-scale benchmark that primarily supports hemorrhage detection and subtype classification on CT, whereas large TBI research platforms such as CENTER-TBI [5] mainly function as extensive clinical research resources with broad longitudinal and structured data collected across multiple centers. In parallel, publicly available pixel-level hemorrhage segmentation datasets remain relatively limited in scale. As a result, there is still a practical gap between the data requirements of end-to-end multimodal severity-assessment pipelines and the datasets that are currently available for fair external validation [6]. This limitation has motivated the use of the matched multimodal version of the 103_TBI cohort in the present study.

Intracranial hemorrhage (ICH) detection is central to TBI severity assessment and primarily relies on brain computed tomography (CT). Accurate ICH volume estimation supports hematoma progression and mortality prediction. Non-contrast CT (NCCT) remains the clinical standard. Supervised deep learning models have shown strong segmentation performance, notably V-Net [7], SegResNet [8], UNETR [9], SwinUNETR [10], and U-Net [11]. A key practical challenge in clinical TBI datasets is the high variability in the number of acquired axial slices across patient volumes (typically ranging from 30 to over 100),

p-ISSN 3093-3285
e-ISSN 3093-3315
https://doi.org/10.51316/jst.192.ssad.2026.36.3.x
Received: Feb 28, 2026; Revised: Apr 3, 2026;
Accepted: May 5, 2026; Online: Jun 7, 2026

depending on scanner protocol and patient positioning. Volumetric 3D models require padding or resampling to a fixed depth, introducing through-plane interpolation artifacts [12, 13] that degrade fine hemorrhagic boundaries. In our experiments, nnU-Net [14] in its 2D slice-based configuration yielded the strongest segmentation performance among all evaluated models by naturally accommodating variable-depth volumes without resampling. Nevertheless, applying 2D models independently per slice leads to class-ID inconsistencies across adjacent slices—slices at the same anatomical location may receive conflicting hemorrhage-type labels. To address this, we propose an anatomy-guided post-processing strategy: the structurally stable calvarium boundary is used to anchor a calvarium-centered polar coordinate transformation [15], converting each variable-geometry axial slice into a fixed-size rectangular representation. Slices are automatically grouped into three anatomically coherent clusters by their relative slice-index, and class labels are assigned via probabilistic projection of training annotations onto the polar coordinate grid, effectively correcting class-ID errors introduced by independent 2D inference.

To this end, this study utilizes a framework on TBI multimodal data, which comprise CT imaging scan data and clinical data in order to assess the severity of TBI. The proposed framework examine various ML algorithms (i.e., LR, NB, kNN, SVM, DT, RF, XGBoost). Among the numerous multimodal information, the most important features such as image-based one have been extracted automatically. This paper's contributions are threefold:

1) A slice-based ICH segmentation framework using nnU-Net [14] that handles variable axial slice counts without volumetric resampling, combined with an anatomy-guided post-processing pipeline comprising: (i) calvarium-centered polar coordinate normalization for registration-free spatial encoding across volumes [15], (ii) automatic slice clustering into three anatomically coherent groups by relative slice-index, and (iii) probabilistic class-ID assignment via projection of training annotations onto the polar grid, correcting cross-slice label inconsistencies and recovering missed peripheral lesion regions.
2) A multimodal framework has been proposed where imaging data from CT brain and clinical information are integrated. By jointly leveraging radiological features and patient-specific clinical information. Experimental results demonstrate that this multimodal integration substantially enhances predictive performance, with the proposed framework achieving an accuracy of 94.20%.
3) From the integrated set of multimodal features, the most informative characteristics of the dataset have been revealed. This contribution enhances the interpretability of the proposed framework by identifying clinically meaningful variables, rather than focusing solely on predictive accuracy. In addition, it provides insights into the relative importance of imaging- and clinical-derived features, facilitating a more transparent decision-making process and supporting the potential clinical applicability of the framework.

2. Related Works

2.1. Image-based TBI Assessment

There are some different approaches for assessing the severity of TBI. Regarding image-based approaches, deep learning on head CT enables automatic detection and segmentation of acute TBI lesions, producing quantitative biomarkers such as lesion volumes and midline shift [16]. Modern models segment multiple lesion types (e.g., epidural, subdural, intraparenchymal) and support midline shift quantification for objective burden assessment [17]. Frameworks like BLAST-CT and subsequent pipelines show that these measurements correlate with clinical severity and therapeutic needs [18]. However, most imaging pipelines remain constrained by single-center training, limited external validation, and variability in acquisition protocols. Moreover, many studies report correlations without closing the loop to end-to-end severity scoring integrated with routine clinical variables.

2.2. Tabular-based and Hybrid Approaches

In the literature, some research works focused on tabular-based approaches. Recent ML methods on tabular data (random forests, gradient boosting, neural networks) often outperform logistic regression and consistently identify lower admission GCS, older age, and metabolic derangements as key predictors [19]. However, predicting full-spectrum outcomes and achieving external generalization remain challenging [20]. These tabular models frequently underuse quantitative imaging biomarkers or encode imaging as coarse categorical variables, limiting sensitivity to structural injury burden. The handling of missingness and class imbalance is also inconsistent between studies, complicating a fair comparison.

Hybrid approaches combine imaging biomarkers with clinical features and improve prediction versus clinical-only models [21]. For example, fusion models integrating CT with routine clinical data outperform traditional scores for mortality prediction [22], and combining CT radiomics with clinical variables improves 6-month outcome prediction and hematoma expansion [21]. Recent reviews also highlight reproducibility and external validation gaps, underscoring the need for careful evaluation [23]. These results motivate multimodal fusion for robust prognostication. Nevertheless, fusion strategies vary widely (early vs. late fusion, feature engineering

vs. end-to-end learning), and ablation protocols are often insufficient to isolate the marginal value of specific imaging biomarkers (e.g., midline shift vs. lesion volumes), limiting clinical interpretability and deployment readiness.

2.3. Slice-based versus Volumetric Segmentation

Two dominant paradigms exist for brain NCCT segmentation: volumetric (3D) and slice-based (2D) approaches. Volumetric models such as V-Net [7], SegResNet [8], UNETR [9], and SwinUNETR [10] operate on the full 3D volume and can leverage inter-slice context to improve anatomical coherence. In practice, however, clinical head CT—especially in TBI workflows—often exhibits substantial variability in the number of axial slices across patients due to scanner protocol, patient positioning, and field-of-view differences. To fit a fixed-depth 3D input, volumes are typically padded or resampled to a common spacing and depth, which can introduce through-plane interpolation artifacts and blur fine boundaries [12, 13]. These issues are particularly undesirable for small or thin hemorrhagic components, where slight blurring can lead to under-segmentation.

In contrast, 2D slice-based methods process each axial slice independently and naturally accommodate variable-depth volumes without enforcing a fixed depth via resampling. While 2D models may lose explicit 3D context, they are memory-efficient and can be competitive in CT segmentation settings where protocol variability is prominent. A widely used baseline is nnU-Net [14], a self-configuring framework that adapts preprocessing, architecture, training, and post-processing to the target dataset. Its 2D configuration has repeatedly been reported as a strong baseline across diverse benchmarks [14], motivating its use as the segmentation backbone in our pipeline.

2.4. Registration-based Normalization and Its Limitations

A natural strategy for reducing inter-patient variability is image registration, aligning scans to a common atlas or template using affine and deformable transformations [24]. Atlas-based methods are a longstanding paradigm in neuroimaging (particularly MRI) and remain attractive as a way to impose anatomical correspondence. However, registration-based normalization faces substantial challenges in the TBI setting.

First, space-occupying lesions, mass effect, and midline shift violate the smoothness and topology assumptions commonly used by deformable registration, often yielding implausible warps near pathology. The registration literature explicitly recognizes the difficulty of capturing tumor-induced mass effect with standard deformable methods [25]. Analogous issues arise in acute TBI where hemorrhage and edema distort normal anatomy, making atlas alignment ill-posed in the most clinically relevant regions. Second, when volumes differ in slice count, through-plane resampling is required before registration; this step can blur edges and reduce contrast under z-axial partial volume effects [12, 13]. Finally, modern deformable registration pipelines add computational overhead that may be undesirable for time-sensitive clinical workflows [24]. These limitations motivate alternative anatomy-aware normalization strategies that avoid deformable registration.

2.5. Coordinate Normalization via Polar Transform

Polar and log-polar representations have a long history in achieving rotation-invariant or shape-normalized image descriptions by mapping content from Cartesian to a radial-angular domain [26]. In medical imaging, such transforms can be anchored to stable anatomical landmarks to reduce variability in pose and apparent geometry. In head CT, the calvarium provides a structurally consistent boundary across patients and is frequently exploited for symmetry analysis and midline-related measurements [27]. Recent work continues to refine calvarium-based midline approximation for robust symmetry analysis [15]. Building on this observation, a polar transform centered at a calvarium-derived reference—the center of a circumscribed circle fitted to the calvarium boundary on each axial slice—converts a variable-geometry cross-section into a fixed-size rectangular representation, providing registration-free coordinate normalization independent of variable slice counts.

2.6. Post-processing for Segmentation Consistency

Even with strong voxel-level accuracy, deep segmentation models can produce anatomically inconsistent outputs, including label fragmentation, spurious detections, and local class confusion. Post-processing steps such as connected-component filtering, morphological operations, and conditional random fields (CRF) [28] are commonly applied to enforce spatial coherence [29]. For slice-wise 2D models applied independently per slice, an additional failure mode is cross-slice label inconsistency for the same physical lesion—often observed as slice-to-slice "label flickering" in multi-class settings. Such errors can be mitigated by incorporating anatomical priors that restrict implausible class co-occurrence and by enforcing inter-slice agreement via volumetric regularization of predictions. Finally, intensity-guided expansion around confirmed hemorrhagic segments can recover missed peripheral lesion regions: classic seeded region growing [30] and superpixel-based expansions such as SLIC [31] provide practical mechanisms to propagate boundaries to adjacent gray-level-consistent regions, complementing neural predictions in low-contrast or partial-volume-affected areas.

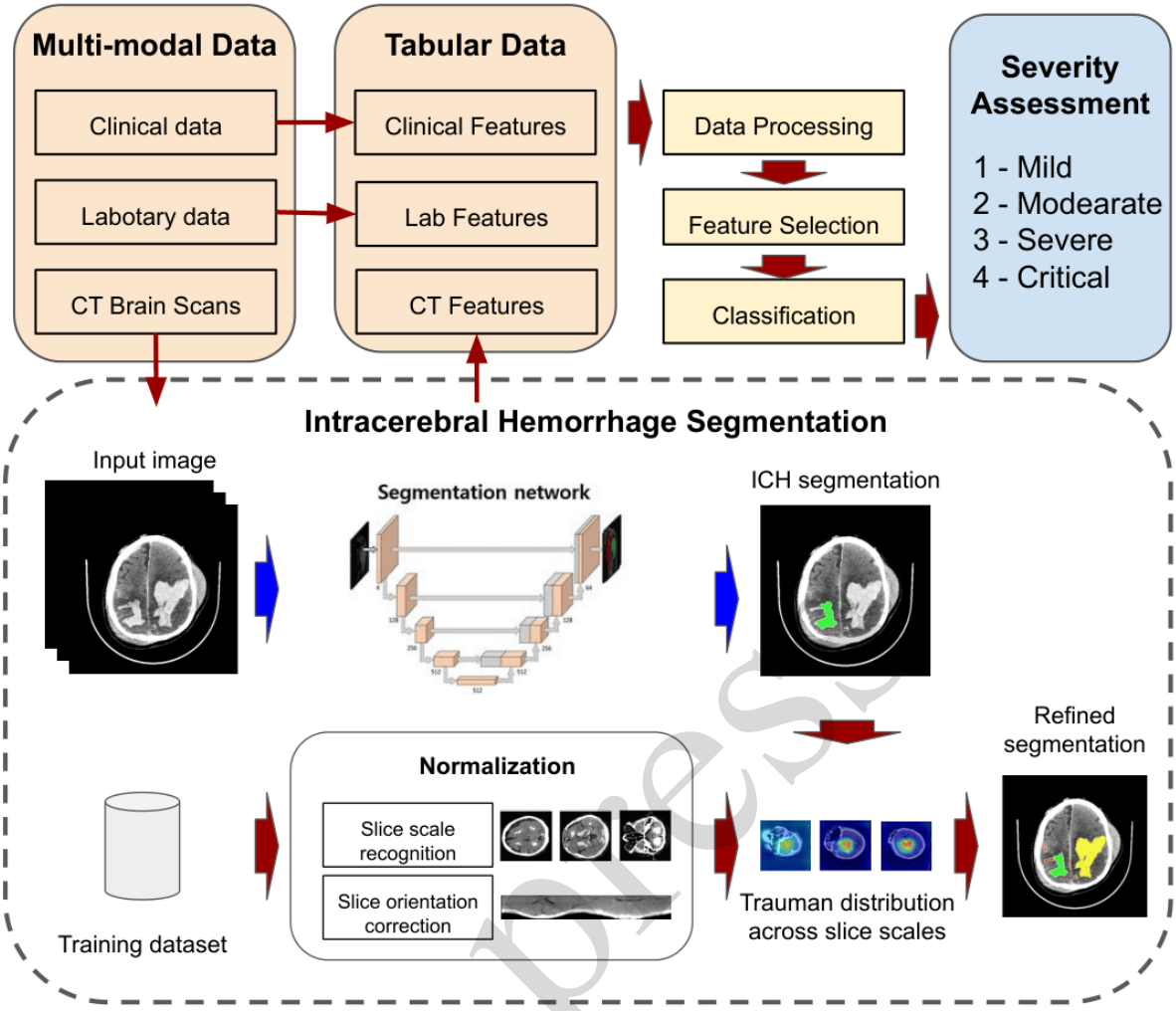

Fig. 1. Overview of the proposed pipeline: Step 1 — Image segmentation and CT feature extraction; Step 2 — Data processing; Step 3 — Classification with Feature Selection and Classification

3. The proposed method

Our proposed method consists of three main parts, shown in Fig. 1. Multimodal data, including data extracted from medical records and CT images, is first processed by Step 1 (image segmentation and CT feature extraction) to produce tabular imaging biomarkers. This data is then processed in Step 2 (data processing) and fed into Step 3 (classification), which includes feature selection (3.1) and classification (3.2).

3.1. Intracerebral Hemorrhage Segmentation in Brain NCCT Scans

3.1.1. Slice-based segmentation with nnU-Net

Clinical TBI head CT volumes exhibit substantial variability in the number of acquired axial slices—typically ranging from 30 to over 100—depending on scanner protocol, patient positioning, and field of view. Volumetric 3D models require padding or resampling to a fixed depth, introducing through-plane interpolation artifacts [12, 13] that are particularly detrimental for thin hemorrhagic regions such as traumatic subarachnoid hemorrhage. To avoid this, we adopt a slice-based (2D) approach in which each axial slice is processed independently, naturally accommodating variable-depth volumes without resampling.

We use nnU-Net [14] in its 2D configuration as the segmentation backbone. nnU-Net is a self-configuring framework that automatically adapts its preprocessing, network architecture, training scheme, and post-processing to the target dataset. In our comparative experiments (Table. 1), the 2D nnU-Net configuration yielded the highest segmentation performance among all evaluated models, confirming the suitability of slice-based processing for this dataset.

3.1.2. Polar coordinate normalization and slice clustering

While 2D slice-based inference avoids resampling artifacts, it introduces a new challenge: independently processed slices may produce inconsistent class-ID assignments for the same physical lesion across adjacent

slices—the same anatomical location may receive different hemorrhage-type labels on consecutive slices. To correct these inconsistencies, we leverage the structural consistency of the calvarium as an anatomical anchor [15].

For each axial slice, the outer boundary of the calvarium is extracted by thresholding at high Hounsfield unit (HU) values characteristic of cortical bone. The center of the minimum circumscribed circle fitted to this boundary defines the origin for a polar coordinate transformation, mapping the slice from Cartesian coordinates (x, y) to polar coordinates (r, θ):

$$r = \sqrt{(x-x_c)^2 + (y-y_c)^2}, \quad \theta = \arctan\left(\frac{y-y_c}{x-x_c}\right), \quad (1)$$

where (x_c, y_c) is the circumscribed circle center. This produces a fixed-size rectangular representation that is consistent across patients and independent of variable slice counts or head positioning, without requiring deformable registration.

Slices across all volumes are then automatically grouped into three clusters using the relative slice-index $s_{\text{rel}} = s/S_{\text{total}} \in [0, 1]$, where s is the absolute slice index and S_{total} is the total number of slices in the volume. Slices at similar relative positions share analogous anatomical cross-sections—superior, mid, and inferior brain levels—providing a principled basis for within-cluster label statistics.

3.1.3. Probabilistic class-ID assignment

Within each cluster, the annotated training labels are projected onto the polar coordinate grid to build per-class probability maps. For each polar cell (r_i, θ_j), the empirical probability of each hemorrhage class c is estimated from the training set:

$$P(c \mid r_i, \theta_j) = \frac{\sum_n \mathbf{1}[\text{label}(r_i, \theta_j, n) = c]}{N_{\text{cluster}}}, \quad (2)$$

where N_{cluster} is the number of training slices in the cluster. During inference, candidate hemorrhagic regions produced by nnU-Net with ambiguous or inconsistent class-IDs are reassigned by selecting the class with the highest probability at the corresponding polar location. Additionally, regions of comparable gray-level intensity adjacent to confirmed hemorrhagic segments are examined for potential extension; the polar probability maps determine whether to assign a class label to newly included voxels. This procedure corrects class-ID errors introduced by independent 2D inference while recovering missed peripheral lesion areas.

As illustrated in Fig. 1, the training dataset is first processed through the normalization stage, including slice-scale recognition and slice-orientation correction, to map annotated lesions into a comparable calvarium-centered space. Within each slice scale, the normalized annotations are aggregated to form a probabilistic trauma distribution that represents the spatial likelihood of each hemorrhage class. During inference, this distribution is used to refine the raw segmentation by reassigning ambiguous lesion regions to the most probable class according to their normalized spatial position.

3.2. Data Processing

After obtaining segmentation outputs, we aggregate imaging biomarkers with structured variables and apply: (1) clinically informed binning for key continuous variables, (2) k-nearest neighbours imputation to retain samples with missing fields, (3) Z-score scaling for numerical stability, and (4) SMOTE over-sampling to mitigate class imbalance in the training data.

Data binning: Many clinical features (e.g., blood pressure, selected laboratory values) are continuous; discretizing them into medically interpretable ranges improves interpretability and downstream explanation. Thresholds follow established clinical guidance (e.g., systolic blood pressure: 90–120 mmHg normal, 121–139 mmHg prehypertension, >140 mmHg hypertension).

Data imputation: We use k-nearest neighbors imputation with $k = 3$. Distances over mixed-type features follow the Heterogeneous Euclidean-Overlap Metric (HEOM), adapted to handle missingness without biasing neighborhood structure.

Data scaling: To prevent domination by high-variance variables, we apply Z-score standardization, which transforms each feature to zero mean and unit variance:

$$\tilde{x} = \frac{x - \mu(x)}{\sigma(x)}. \quad (3)$$

Data over-sampling: We employ SMOTE [32–34] to synthesize minority-class samples by interpolating among k-nearest neighbors, improving balance without simple duplication.

3.3. Classification Models

This section outlines the classification workflow with two stages: feature selection and predictive modeling. We first identify key features using interpretable methods, then evaluate multiple ML classifiers to assess predictive value in TBI outcomes.

3.3.1. Feature Selection

Feature selection reduces complexity, prevents overfitting, and enhances interpretability, which is crucial in clinical contexts. Using all features may obscure patterns and increase computational cost, whereas medical diagnosis often depends on a limited set of indicators. We applied three representative techniques: Mutual Information (MI, Filter), SHAP

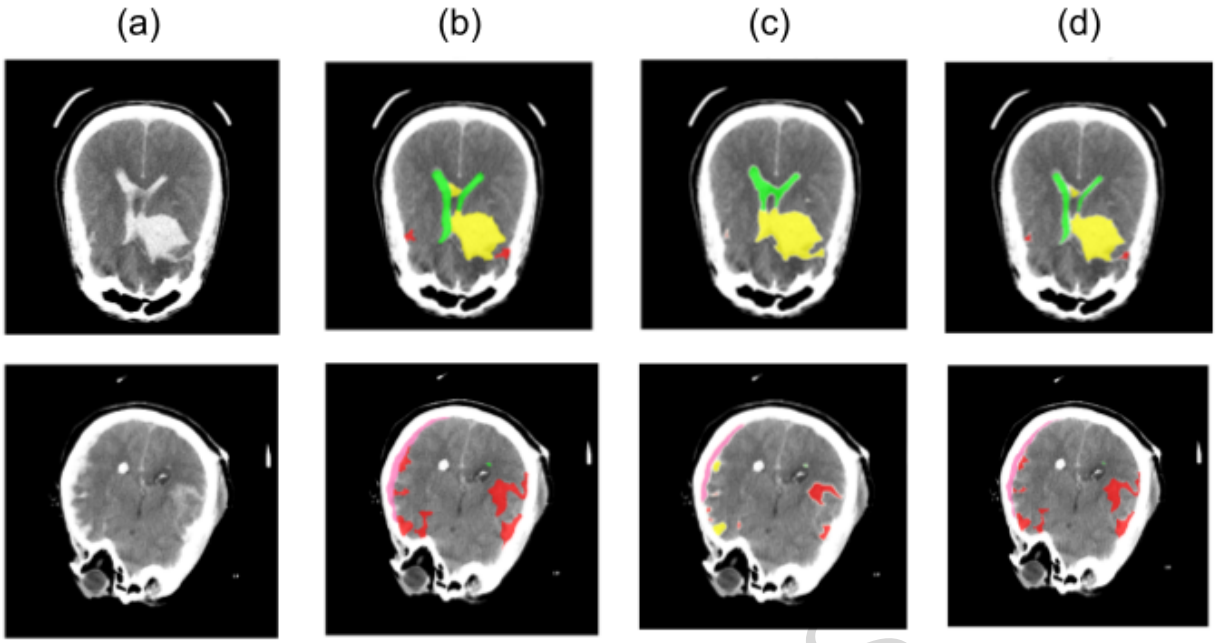

Fig. 2. Illustration of the proposed segmentation pipeline. (a): NCCT axial slice. (b): Ground truth overlay. (c): Raw nnU-Net 2D prediction (note class-ID inconsistencies). (d): Refined segmentation after polar-based class-ID correction. EDH IPH IVH SAH SDH

(Embedded), and DiCE (Wrapper). Each highlights feature relevance from different perspectives: MI provides model-independent statistics, SHAP derives attribution from trained models, and DiCE captures causal influence via counterfactuals.

For consistency, we retained the top 17 features per method, based on a prior MRMR threshold of 0.5. In addition, clinical experts identified 34 medically significant features; overlap with algorithmic results served as secondary validation.

3.3.2. Classification models

We implemented six ML algorithms across diverse paradigms for benchmarking:

- **Linear:** Logistic Regression (LR), Naive Bayes (NB).
- **Distance-Based:** k-Nearest Neighbors (KNN).
- **Kernel-Based:** Support Vector Machine (SVM).
- **Tree Ensembles:** Decision Tree (DT), Random Forest (RF), XGBoost (XGB).

This heterogeneous set enables robust evaluation, from probabilistic to tree-based methods. Notably, RF and XGB are well-suited for structured medical data, offering feature importance and handling nonlinearities. Together, these models provide a solid baseline for assessing feature selection strategies and prediction of TBI outcomes.

4. Experiments

4.1. Dataset

We used the single-center 103_TBI dataset comprising 504 anonymized adult patient records with corresponding head CT scans collected at 103 General Hospital (Hanoi, Vietnam). Each record includes structured clinical variables (vital signs, neurologic examination, accident context) and laboratory tests, alongside CT-derived findings. The target label is a 4-level clinical severity score (1: mild to 4: critical) determined by expert consensus. Severity level 2 is the majority class; class imbalance is addressed during training via SMOTE. Although the overall cohort comprised 504 patients, the complete multimodal pipeline was evaluated on the subset of 68 patients for whom both structured clinical/tabular data and CT imaging data with lesion annotations were available. This matched subset provided the required inputs for end-to-end assessment of the proposed framework.

4.2. Experimental Settings

To evaluate model performance in a robust and reproducible manner, we designed an experimental protocol that includes stratified data splitting, cross-validation, hyperparameter optimization, and statistical averaging over multiple trials.

To evaluate the proposed framework in a clear and reproducible manner, the experiments have been organized into two complementary parts according to the available data configuration in the 103_TBI cohort. First, experiments on the full cohort of 504 patients followed the general classification protocol described

above, including stratified train–test splitting, cross-validation, hyperparameter optimization, and repeated runs across multiple random seeds. This part was intended to examine classification performance on the broader clinical dataset and to identify the most informative features for TBI severity assessment. The resulting analysis served to determine the key variables contributing to the prediction task.

Second, a matched subset of 68 patients was extracted from the same cohort, including only cases with complete structured clinical/tabular data and CT imaging data accompanied by lesion annotations. This subset supported end-to-end evaluation of the proposed multimodal pipeline. After segmentation of the different lesion types on CT, the lesion burdens were quantitatively measured to derive imaging biomarkers. These imaging-derived quantities were then used to replace the corresponding lesion-related tabular attributes, yielding an updated multimodal representation for each patient. Classification experiments on this subset were finally performed to assess the effectiveness of the segmentation-driven features in supporting TBI severity prediction.

Data splitting and evaluation metrics. Prior to training, the dataset was split into training (80%) and hold-out test (20%) sets using stratified sampling to preserve class distribution. For training, a 10-fold cross-validation (CV) scheme was applied to mitigate data partition bias. Each fold produced a model, and the best-performing estimator (based on validation performance) was selected and retrained on the full training set before final evaluation on the hold-out test set. Model performance was assessed using standard classification metrics: Accuracy, Precision, Recall, and F1-score. To ensure statistical robustness, the entire pipeline was repeated across 50 random seeds, and the final scores were reported as the mean of all trials.

Hyperparameter tuning. We applied both grid search and randomized search strategies to optimize model parameters. In grid search, each hyperparameter was varied over a predefined range. Randomized search was employed to explore a broader configuration space efficiently. Hyperparameter tuning was embedded within the cross-validation loop to prevent data leakage into the test set.

4.3. CT-Brain Imaging Segmentation

Figure 3 presents representative training curves of the evaluated segmentation models. The 2D nnU-Net is included as the principal baseline in this study. Consistent with the quantitative results in Table 1, the training curves in Fig. 3 show that the 2D nnU-Net achieves the strongest baseline performance among the compared architectures, further supporting the suitability of slice-based processing for clinical CT data with inconsistent slice depth. The quantitative evaluation is shown in Table 1. The values in the HD95 column denote the 95th percentile Hausdorff distance (lower is better). Our proposed method (the combination of nnU-Net 2D and polar-based class-ID correction) further improves both Dice and IoU over the nnU-Net (2D) baseline.

Table 1. Comparison of segmentation performance. **Ours** denotes nnU-Net (2D) with polar-based class-ID correction and region extension.

Method	**Dice (%)**	**IoU (%)**	**HD95**
V-Net	62.69	51.30	52.08
SegResNet	55.32	42.27	101.13
UNETR	54.86	41.72	136.05
SwinUNETR	52.03	39.20	169.30
U-Net (3D)	64.19	52.11	55.07
Ours	**67.21**	**55.23**	54.17

Although the proposed method achieves the best segmentation performance among the compared models, the obtained Dice score still indicates that precise lesion recognition remains challenging in traumatic brain CT. Hemorrhagic lesions often exhibit irregular shapes, heterogeneous density, and indistinct boundaries, which make accurate voxel-level segmentation difficult, particularly near peripheral or low-contrast regions.

The relatively weaker performance of the 3D models can be explained mainly by the substantial variation in the number of axial slices across patient volumes in this dataset. In order to train 3D architectures, the input volumes must be normalized to a more consistent depth, typically through padding or resampling. This volumetric standardization may alter inter-slice anatomical continuity and reduce the representation of fine lesion structures, especially for small or thin hemorrhagic components. In contrast, the 2D nnU-Net processes each slice directly and therefore avoids the need for fixed-depth normalization, making it better suited to the inconsistent depth characteristics of the present clinical CT dataset. In addition, 3D models require heavier preprocessing and substantially higher computational cost, which further limits their practical effectiveness under this data setting.

From a deployment perspective, computational efficiency is an important consideration for integration into emergency workflows. In our current implementation, inference was performed on an NVIDIA A100 GPU, with an average processing time of approximately 2.5 seconds per CT volume. Although emergency applications require rapid analysis, the proposed framework is intended as a time-sensitive clinical decision-support tool rather than a strict real-time system.

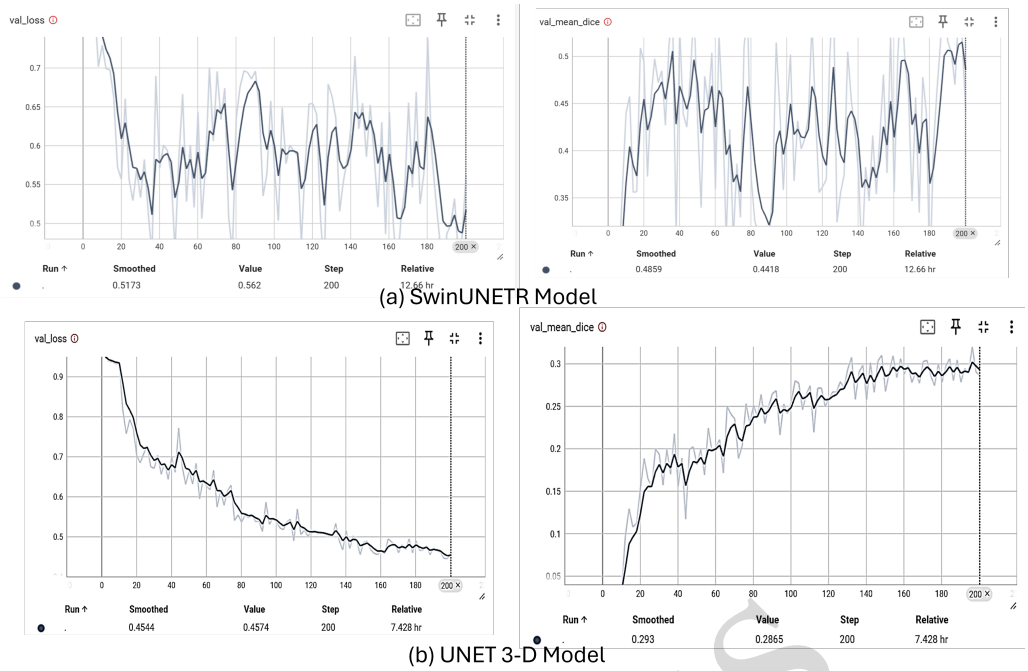

Fig. 3. Validation performance metrics during training of CT-brain two segmentation models. (a): SwinUNETR Model;(b): UNet-3D model; in each row: left panel is training loss, and right panel is Mean Dice coefficient

4.4. Informative features for the TBI severity assessment

To reduce redundancy and improve interpretability, we ranked features using three complementary methods: Mutual Information (filter), SHAP attributions (embedded), and DiCE-based counterfactual change rates (wrapper). Following prior MRMR-based sensitivity checks, we retained a compact subset (e.g., top-17) for model training. As this work emphasizes imaging-derived biomarkers, detailed rankings and overlaps are omitted for brevity.

Table 2 summarizes the key variables identified by the feature analysis as the most influential factors associated with TBI severity and subsequently used in the classification stage of the proposed framework.. These influential features are used as inputs to predict TBI severity, thereby providing a clearer view of the model logic. The results show that, in addition to core clinical indicators such as GCS and vital signs, CT-related variables also play an important role, particularly those derived from the segmentation and quantification of hemorrhagic lesions. This finding highlights the added value of CT imaging biomarkers, including segmentation-derived measures, in improving severity assessment.

4.5. Classification Results

Table 3 summarizes the classifiers' performance under different search and optimization strategies. GridS, RandS, and BayesO denote Grid Search, Random Search, and Bayesian Optimization, respectively; LOOCV indicates Leave-One-Out Cross-Validation;

Table 2. Top influential features identified by the proposed framework for TBI severity assessment.

Rank	Feature
1	Glasgow Coma Scale (GCS)
2	Midline shift width
3	Intraparenchymal hemorrhage-related measure
4	Subdural hematoma thickness / volume
5	Epidural hematoma volume
6	Breathing rate
7	Pulse
8	Time from accident to hospital
9	Pupillary response / pupil size
10	Basal cistern or Rotterdam-related CT feature

SMT marks SMOTE usage. Bolded values highlight the best model per configuration, and underlined ones the best configuration per model. On the 103_TBI dataset, tree-based ensembles dominated: XGBoost reached 94.20% accuracy (Random Search + SMOTE), followed by Random Forest (93.92%). SMOTE consistently improved accuracy by 3–6%, while stochastic searches (RandS, BayesO) often outperformed GridS.

Further analysis compared XGBoost, Random Forest, Decision Tree, and kNN across varying feature subsets and five metrics (Accuracy, Precision, Recall, F1-score, AUC). XGBoost showed the highest overall performance, peaking near 17 features (accuracy $\approx$93.6%) before slightly declining beyond $\sim$30 features due to redundancy. Random Forest improved with larger subsets, reflecting bagging robustness. Decision Tree remained stable under depth = 5, while kNN was sensitive to dimensionality—best with 5–10 features but degrading as feature count increased.

Table 3. Cross-validation and model performance on the 103_TBI dataset

Search	CV	SMT	LR	SVM	DT	kNN	RF	XGB
	LOOCV		85.12	84.34	85.21	77.14	**88.50**	**88.50**
	LOOCV	x	84.70	84.15	87.78	80.57	**93.38**	93.08
	10-fold		84.05	83.15	87.90	76.99	87.50	**89.50**
	10-fold	x	84.72	84.71	88.88	78.89	92.51	**94.17**
GridS	10-fold		86.22	86.50	89.32	77.49	88.49	**88.99**
GridS	10-fold	x	84.41	85.79	91.13	82.80	**93.92**	93.65
RandS	10-fold		84.50	86.22	**89.49**	77.49	88.99	88.99
RandS	10-fold	x	81.41	93.92	90.84	82.80	**94.20**	93.65
BayesO	10-fold		84.50	84.93	**89.99**	77.49	88.49	**88.99**
BayesO	10-fold	x	84.39	**93.65**	91.41	82.80	**93.92**	93.65

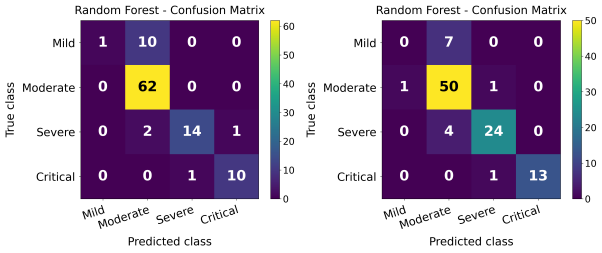

Fig. 4. Representative confusion matrices of the Random Forest classifier for the four-level TBI severity classification task, shown with SMOTE and without SMOTE

To provide a more detailed view of class-wise performance under the imbalanced severity distribution, Fig. 4 presents representative confusion matrices of the Random Forest classifier before and after applying SMOTE. The overall confusion pattern remains similar in both settings, suggesting that the use of SMOTE does not substantially alter the overall accuracy trend. In particular, the critical class is identified relatively well, while some confusion between mild and moderate cases remains. These results indicate that SMOTE improves class balance without markedly distorting the class-wise prediction performance.

5. Conclusion

This study presents an imaging-centric, interpretable machine learning framework for classifying traumatic brain injury (TBI) severity using the single-center 103_TBI dataset of 504 patients. We adopt a slice-based segmentation pipeline using nnU-Net, which avoids volumetric resampling artifacts arising from variable slice counts in clinical CT data. To address class-ID inconsistencies across adjacent slices—a key failure mode of independent 2D inference—we propose an anatomy-guided post-processing strategy grounded in calvarium-centered polar coordinate normalization. Slices are automatically clustered into three anatomically coherent groups by their relative slice-index, and class labels are assigned via probabilistic projection of training annotations onto the polar grid. The quantitative biomarkers extracted (lesion volumes, midline shift, subarachnoid characteristics), when integrated with core clinical variables, deliver robust performance with XGBoost and Random Forest (up to 94.20% accuracy). Consequently, the systematic pre-processing, class rebalancing with SMOTE, and selected feature subsets are important for reliable performance. These results are added value of segmentation-derived imaging biomarkers for TBI's assessment and their suitability for digital health integration.

Future work: The generalizability of the proposed framework beyond the current dataset should be interpreted with caution, as its performance may be influenced by differences in patient populations, acquisition protocols, annotation practices, and the availability of matched multimodal data. Future work will therefore focus on validation on larger multicenter cohorts, further refinement of the polar normalization and probabilistic class-ID assignment, and integration of the pipeline into time-sensitive clinical decision-support systems.

Acknowledgments

This research is funded by the Ministry of Education and Training (MOET) under grant number B2023-BKA-09 "Research and development of supporting tools for prognosis of traumatic brain injury using multimodal information and artificial intelligence."

Reference

[1] E. National Academies of Sciences, Medicine, Health, M. Division, B. on Health Care Services, B. on Health Sciences Policy, C. on Accelerating Progress in Traumatic Brain Injury Research, and Care, Traumatic Brain Injury: A Roadmap for Accelerating Progress, National Academies Press, 2022.

[2] D. Agrawal, S. Joshi, and L. Poonamallee, Automated midline shift detection and quantification in traumatic brain injury: A comprehensive review, Indian Journal of Neurotrauma, vol. 21, no. 01, pp. 006–012, 2024.

[3] Q. Feng, Y. Zhao, and J. Wang, Application of machine learning techniques in predicting outcomes for traumatic brain injury: A multicenter study, Scientific Reports, vol. 13, pp. 28188, 2023. https://doi.org/10.1038/s41598-023-28188-w.

[4] A. E. Flanders, L. M. Prevedello, G. Shih, S. S. Halabi, J. Kalpathy-Cramer, R. Ball, J. Mongan, A. Stein, F. C. Kitamura, M. P. Lungren, *et al.*, Construction of a Machine Learning Dataset through Collaboration: The RSNA 2019 Brain CT Hemorrhage Challenge, Radiology: Artificial Intelligence, vol. 2, iss. 3, pp. e190211, 2020. https://doi.org/10.1148/ryai.2020190211.

[5] CENTER-TBI, Data Access & Publication Requests, 2025. [Online] https://www.center-tbi.eu/data.

[6] A. Spahr, D. Frey, M. Béranger, R. Wiest, and M. Reyes, Label-Efficient Deep Semantic Segmentation of Intracranial Hemorrhages in Head CT, Frontiers in Neuroimaging, vol. 2, pp. 1157565, 2023. https://doi.org/10.3389/fnimg.2023.1157565.

[7] F. Milletari, N. Navab, and S.-A. Ahmadi, V-net: Fully convolutional neural networks for volumetric medical image segmentation, pp. 565–571, 2016.

[8] A. Myronenko, 3D MRI brain tumor segmentation using autoencoder regularization, pp. 311–320, 2018.

[9] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu, Unetr: Transformers for 3d medical image segmentation, pp. 574–584, 2022.

[10] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu, Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images, pp. 272–284, 2021.

[11] O. Ronneberger, P. Fischer, and T. Brox, U-net: Convolutional networks for biomedical image segmentation, pp. 234–241, 2015.

[12] P. Monnin, N. Sfameni, A. Gianoli, and S. Ding, Optimal slice thickness for object detection with longitudinal partial volume effects in computed tomography, Journal of Applied Clinical Medical Physics, vol. 18, iss. 1, pp. 251–259, 2017. https://doi.org/10.1002/acm2.12005.

[13] S. Gaj, P. Singh, S. Panigrahi, and J. Sivaswamy, Towards texture accurate slice interpolation of medical images using PixelMiner, Computers in Biology and Medicine, vol. 156, pp. 106701, 2023. https://doi.org/10.1016/j.compbiomed.2023.106701.

[14] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation, Nature Methods, vol. 18, iss. 2, pp. 203–211, 2021. https://doi.org/10.1038/s41592-020-01008-z.

[15] A. Nogueira-Rodríguez, L. Rebourda, F. Giraldez, *et al.*, Brain Midline Approximation to Improve Symmetry Analysis of Brain CT Scans, Medical Engineering & Physics, vol. 135, pp. 104285, 2025. https://doi.org/10.1016/j.medengphy.2024.104285, Available online late 2024, official issue 2025.

[16] P. Rajpurkar, E. Chen, O. Banerjee, and E. J. Topol, AI in health and medicine, Nature Medicine, vol. 28, iss. 1, pp. 31–38, 2022.

[17] J. Brown, B. Cheng, C. Yao, *et al.*, BLAST-CT: A deep learning framework for automated segmentation and classification of traumatic brain injury lesions on CT, NeuroImage: Clinical, vol. 35, pp. 103120, 2022.

[18] P. Brossard *et al.*, AI-based lesion segmentation on admission CT predicts therapeutic intensity in severe traumatic brain injury, Critical Care, vol. 27, iss. 1, pp. 83, 2023.

[19] A. Marino *et al.*, Systematic review and meta-analysis of machine learning models for traumatic brain injury outcome prediction, Frontiers in Neurology, vol. 15, pp. 131245, 2024.

[20] D. Bark *et al.*, Predicting full-spectrum functional outcomes after traumatic brain injury with machine learning, Journal of Neurotrauma, vol. 41, iss. 5, pp. 765–777, 2024.

[21] L. Zhang *et al.*, Radiomics-based machine learning model for outcome prediction in traumatic brain injury, European Radiology, vol. 32, pp. 2306–2316, 2022.

[22] M. Pease *et al.*, Fusion of head CT and clinical data for improved outcome prediction in severe traumatic brain injury, Frontiers in Neurology, vol. 13, pp. 877341, 2022.

[23] D. Rojas *et al.*, Artificial intelligence for outcome prediction in traumatic brain injury: a systematic review of clinical applications, Neurosurgical Review, vol. 48, iss. 2, pp. 229–241, 2025.

[24] B. B. Avants, N. J. Tustison, G. Song, P. A. Cook, A. Klein, and J. C. Gee, A reproducible evaluation of ANTs similarity metric performance in brain image registration, NeuroImage, vol. 54, iss. 3, pp. 2033–2044, 2011. https://doi.org/10.1016/j.neuroimage.2010.09.025.

[25] A. Mohamed, E. I. Zacharaki, D. Shen, and C. Davatzikos, Deformable registration of brain tumor images via a statistical model of tumor-induced deformation, Medical Image Analysis, vol. 10, iss. 5, pp. 752–763, 2006. https://doi.org/10.1016/j.media.2006.06.005.

[26] P.-T. Yap, X. Jiang, and A. C. Kot, Two-dimensional polar harmonic transforms for invariant image representation, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, iss. 7, pp. 1259–1270, 2010. https://doi.org/10.1109/TPAMI.2009.119.

[27] R. Liu, S. Li, B. Su, C. L. Tan, T.-Y. Leong, B. C. Pang, C. C. T. Lim, and C. K. Lee, Automatic detection and quantification of brain midline shift using anatomical marker model, Computerized Medical Imaging and Graphics, vol. 38, iss. 1, pp. 1–14, 2014. https://doi.org/10.1016/j.compmedimag.2013.11.001.

[28] P. Krähenbühl and V. Koltun, Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials, Advances in Neural Information Processing Systems (NeurIPS), vol. 24, 2011.

[29] K. Kamnitsas *et al.*, Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation, Medical Image Analysis, vol. 36, pp. 61–78, 2017.

[30] R. Adams and L. Bischof, Seeded Region Growing, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 16, iss. 6, pp. 641–647, 1994. https://doi.org/10.1109/34.295913.

[31] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, SLIC superpixels compared to state-of-the-art superpixel methods, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, iss. 11, pp. 2274–2282, 2012. https://doi.org/10.1109/TPAMI.2012.120.

[32] Y. Zhou, S. Aryal, and M. R. Bouadjenek, A Comprehensive Review of Handling Missing Data: Exploring Special Missing Mechanisms, arXiv preprint arXiv:2404, 2024.

[33] M. S. Santos, P. H. Abreu, A. Fernández, J. Luengo, and J. Santos, The impact of heterogeneous distance functions on missing data imputation and classification performance, Engineering Applications of Artificial Intelligence, vol. 111, pp. 104791, 2022. https://doi.org/10.1016/j.engappai.2022.104791.

[34] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.