Multi-Criteria Evaluation Framework for Deep Learning Architectures in Medical Image Segmentation

Toluwani A. Oyewusi[1], Desmond E. Ighravwe[1,2], Moses O. Babatunde[1,2], Abraham O. Amole[1,3], and Sunday T. Ajayi[2]

[1]Bells AI centre, Bells University of Technology, Ota, Nigeria.
[2]Department of Mechanical Engineering, Bells University of Technology, Ota, Nigeria.
[3]Department of Electrical, Electronic and Telecommunication Engineering, Bells University of Technology, Ota, Nigeria.
*Corresponding author email: ighravwedesmond@gmail.com

Direct Research Journal of Engineering and Information Technology

Vol. 13(3), Pp. 99-110, December 2025,

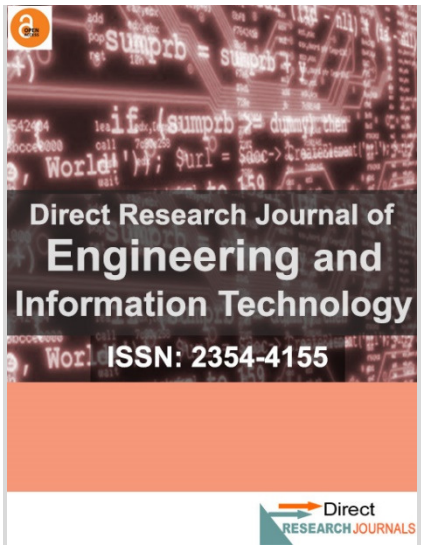
Research Article
ISSN: 2354-4155
https://journals.directresearchpublisher.org/index.php/drjeit; https://www.ajol.info/index.php/drjeit

Received 5 September 2025, Accepted 20 November 2025, Published 11 December 2025

ABSTRACT

The rapid expansion of deep learning architectures for medical image segmentation presents a major challenge to selecting optimal model to implement in practice. While several traditional evaluation methods have been proposed to solving this selection problem, they cannot use criteria interdependencies and expert judgment to select deep learning architectures for medical image segmentation. This study, therefore, presents a multi-stage multi-criteria decision-making framework to address this knowledge gap. The framework contains an improved DEMATEL (Decision Making Trial and Evaluation Laboratory), intuitionistic fuzzy AHP (Analytic Hierarchy Process), and adaptive VIKOR (VIekriterijumsko Optimizacija Kompromisno Resenje). This study evaluated framework performance using six deep learning architectures and thirteen criteria. The DEMATEL-fuzzy AHP results showed sensitivity, specificity, and accuracy the most important criteria for evaluate the deep learning architectures for medical image segmentation. The enhanced VIKOR model identified Swin-Unet and nnU-Net as the optimal compromise and best solution, respectively, for the optimum compromise. Based on sensitivity analysis that was conducted, the VIKOR ranked architecture as UNETR >> nnU-Net >> Attention U-Net >> TransUNet >> Swin-Unet >> classic U-Net. This study's findings have showed that the proposed framework can be used to support deep learning architectures for medical image segmentation decisions.

Keywords: *Deep learning architecture selection, medical image segmentation, multi-criteria decision-making, clinical decision support*



Citation: Oyewusi, T. A., Ighravwe, D. E., Babatunde, M. O., Amole, A. O., & Ajayi, S. T. (2025). Multi-Criteria Evaluation Framework for Deep Learning Architectures in Medical Image Segmentation. *Direct Research Journal of Engineering and Information Technology*, Vol. 13(3), Pp. 99-110.
https://doi.org/10.26765/DRJEIT83396604

INTRODUCTION

Medical imaging methods, such as Magnetic Resonance Imaging (MRI), Computed Tomography (CT), ultrasound and X-ray, generate huge amounts of visual data that require extensive interpretation to diagnose patients (Abhisheka *et al*., 2024). The data have improved the accuracy and efficiency of diagnosing and monitoring disease (Li *et al*., 2023). The methods make it possible to learn complex image segmentation tasks that were done manually by trained radiologists and doctors. These methods are developed using deep learning architectures, including convolutional neural networks or transformers (Ghribi and Hamdaoui, 2025). Due to the existence of several medical imaging methods, medical professionals find it difficult to select the most suitable deep learning architectures in medical image segmentation. This selection problem is a multi-criteria decision making (MCDM) problem because of the need to use several criteria to select suitable deep learning architectures in medical image segmentation.

Classic model selection approaches cannot solve this selection problem because they rely on single-criterion analyses. Also, they cannot account for the complicated trade-offs between a real-world medical application and the complex determinants of models (Angelis, 2017). For example, a model can be better but impractical to implement due to expensive computational resources, not able to interpret the health care condition, or bias towards a demography (Ojha *et al*., 2025). These complex requirements require a systematic, wide-ranging assessment that combines different performance criteria.

While clinical implementation requires more consideration for computational efficiency, model interpretation, robustness to change, generalizability for diverse patient populations, and ethical considerations for algorithmic fairness, clinical deployment are interdependent (Weiner *et al*., 2025).

The development of deep learning models for medical imaging has been mostly focused on performance measures or on both pairwise and architectural comparisons (Suzuki, 20017). A number of benchmark studies have created performance hierarchies by segmentation accuracy measures such as the Dice coefficient or Intersection over Union (IoU) on standard datasets. But, some of these approaches have flaws. First, they often consider evaluation criteria as independent. For example, model interpretability can influence medical adoption rates, which affects the amount of available annotated data for model refinement to be compared and ultimately affects generalizability. Second, conventional evaluation methods often use equal weighting schemes or subjective expert judgments and are not adequately understood in terms of uncertainty of the importance of criterion importance. Third, several existing frameworks do not account for the influence of time in the relationship between spatial variability and system performance during deployment lifecycles. While the implications for deep learning architecture selection for medical imaging remain a challenge, multi-criteria decision-making (MCDM) methods provide theoretical empirical support for dealing with these evaluation issue (Kulak *et al*., 2015). In many healthcare technology assessment settings, MCDM techniques, such as conventional Analytic Hierarchy Process (AHP) and Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS), have been used to generate medical information in the literature. These techniques have limitations in uncertainty, subjective judgments, and dynamic interactions of deep learning model evaluation. In particular, they cannot show expert hesitation, entail non-membership degrees in preference assessments, integrate feedback loops in criteria, and adapt to changing performance relationships over time.

New developments in fuzzy set theory and causal analysis methods have opened the door to new ways of addressing the limitations mentioned above. Intuitionistic fuzzy sets provide a mathematical architecture for explicitly representing membership, non-membership, and hesitation grades in expert judgments. It provides a detail representation of uncertainty than conventional fuzzy approaches. DEMATEL (Decision-Making Trial and Evaluation Laboratory) analyse the causal impact relationships with assessment criteria, uncovering influence mechanisms that traditional approaches fail to capture. Lastly, VIKOR (VlseKriterijumska Optimizacija I Kompromisno Resenje) allows compromise solutions to be identified. These MCDM models have not been used to develop a framework for deep learning architecture assessment in medical imaging.

This study addresses these shortcomings by proposing an adaptive multi-stage MCDM framework. The framework synthesizes DEMATEL with temporal dynamics, intuitionistic fuzzy AHP and adaptive VIKOR. One of the novelties of the framework is that integrates temporal influence decay and feedback amplification into the DEMATEL method to enhance the analysis of criteria within the time-frame when deployed. The framework used triangular intuitionistic fuzzy numbers with explicit hesitation quantification in the AHP pairwise comparison process to address expert uncertainty in criterion weighting. Furthermore, the framework used a time-adaptive integration coefficient, which dynamically balances AHP weights with DEMATEL results to scores as system learning progresses. Also, the framework sets multi-level utility thresholds, ideal, aspiration, tolerable, and anti-ideal to improve standard VIKOR method performance.

Literature review

Several innovative studies have been documented on medical imaging. For example, Otsu introduced a method

that separates the foreground and background by automatically picking a threshold value that maximizes the differences of the pixel values between those groups. Their focus was on the simplest image thresholding which required the grayscale histogram without any prior information about the image. This method followed a simple procedure that has a wide range of applications in image segmentation. Another widely used method involves an iterative process of threshold selection to separate an object from its surrounding. An initial threshold value is chosen to separate the pixels into groups and based on the mean intensities of those groups (object and background), the new threshold is computed as the average of the mean intensities of those two groups. The method is repeated until the threshold remains unchanged. The goal is to sample the object background without including any of the object's pixels or irrelevant background information. The advantage of this iterative approach is that the object is always shown as white on a black background. Thresholding has important applications in text recognition, medical image enhancement and nuclear particle analysis (Ridler and Calvard, 1978). Due to the limitations of Otsu's method, another threshold selection mechanism was proposed. This method, invented by Kittler and Illingworth, uses probability as a basis to determine which threshold produces the lowest probability of pixel misclassification i.e., misclassifying objects and their background (Kittler and Illingworth, 1985)

Maximum entropy thresholding was an improvement on alternative methods due to the inability of those global thresholding methods to make use of the important features of the image (Wong and Sahoo, 1989). This method of thresholding uses the concept of maximum entropy by splitting the pixels in the object and background in the most informative way. Ultimately, the threshold selected is the one that maximizes the entropy or uncertainty associated with the pixel intensities. The result displays the most meaningful variation of the image splits which provide the most information about the image. Global thresholding loses many features in the process of generating binary images. Adaptive thresholding solves this problem of varying illumination in the image by picking different thresholds for each part of the image (Weeks *et al*., 1994).

Canny (1986) identified a computational approach to detecting edges. He explained that edge detectors are vulnerable to several errors and his approach was to design edge detectors with a low error rate, good localization and only one response per edge. The first step is image smoothening to reduce noise, followed by calculating gradients, and then thinning edges for precision. A high and low threshold are used to determine which features correspond to edges. This approach applies to a wide range of images with differing noise levels and edge widths.

To represent the outline of an object, contours connect edges to form boundaries. This idea introduced active contour models called snakes, which are interactive curves directed by internal and external forces to efficiently localize objects. Snakes have important benefits in image segmentation and motion tracking. The traditional approach involved identifying individual image contours and then connecting them together. Snakes use the concept of energy minimization to distort its shape to match the nearest contour. Their dynamic behaviour is caused by constant energy minimization. Essentially, the snakes are looking for the global minimum where the energy is lowest and this represents the edges of the object (Kass, 1988). Osher and Sethian (1988) developed numerical algorithms called Propagation with Speed depending on Curvature (PSC) algorithms. Older methods used marker points on the curve to represent the moving boundary and were extremely accurate for small motions. For more complex motions, the points can get improperly clustered on steep curves and the approach fails to accurately handle when boundaries merge or split. PSC algorithms are for tracking edges moving with a speed that depends on the local curvature of that area. This provides an automatic mathematical approach to handling the problem of evolving boundaries without the need for constant manual intervention. The motion equation used to determine the curvature-dependent speed is an initial-value Hamilton-Jacobi equation. In image segmentation, this algorithm allows for image boundaries to gradually move until they fit around the edges of the image.

The watershed algorithm, detailed by Vincent and Soille, is a very flexible algorithm that can be used in multi-dimensional images and digital grids. This algorithm was shown to be more accurate than its slow and error prone counterparts that existed at the time. This algorithm was implemented with a breadth-first search approach along with a first-in-first-out (FIFO) queue. The pixel values are sorted in ascending order of intensity and then a breadth-first scan with a queue is used to expand the region outwards. In image segmentation, the watershed algorithm is one of the most potent segmentation tools. The simple computation of the watersheds of a gradient can result in over segmentation where the important contours are lost. This can be solved by modifying the gradient and removing irrelevant contours. Image segmentation is done by simulating flooding, starting from the markings of the objects that will be segmented. The pixels i.e water in this scenario, are spread through the breadth-first search algorithm and a boundary is created once pixels from different markings meet (Vincent and Soille, 1999).

Region growing, an algorithm with some similarities to watershed, requires the use of pixels or regions called seeds, to establish areas that represent segmented regions of an image based on the closeness of pixel values (Adams and Bischof, 1994). This process is continued until no more pixels can be merged or the region cannot grow anymore. Another step in the serial region technique that includes region growing, is region merging.

This step involves merging small regions as long as the intensities of those pixels to be merged do not have a difference more than a certain given threshold value. Region growing is very computationally expensive but it is good for handling noise and object occlusion, while also effectively managing the scale of the segmentation (Yu *et al*., 2023).

For many decades now, clustering has been an important tool in image segmentation and its use dates all the way back to the 1960s. Clustering is described as the unsupervised classification of features into groups called clusters. In image segmentation, the features are the image pixels and intensity while the clusters refer to the image segments (Jain *et al*., 1999). A more efficient version of the K-means clustering algorithm was introduced due to the high computational cost when applied on large amounts of data. The aim of this algorithm is to minimize the within-cluster sum of squares (WCSS) of each cluster i.e. the distance of each point in the cluster from the centre of the cluster. The old version of the algorithm would require a check of all possible cluster combinations. Instead of computing all possible groupings, the improved version looks for a local optimum, where moving any points between clusters would increase the WCSS (Hartigan and Wong, 1979).

Schachter *et al*. (1979) proposed a way to segment images using features like the intensity and location of each pixel rather than the standard clustering method. This approach is useful in complex images because the segmentations were more accurate and connected regions were segmented as a whole. Agglomerative clustering has been applied in unsupervised learning of clusters i.e., segments, in images. This algorithm uses the coefficients of the polynomial curve that represents the pixels of the image. The similarities between the coefficient vectors are calculated using the Mahalanobis distance and the produced result determines which pixels are clustered together. For textured images which are not smooth, the polynomial model was not suitable, so a Markov Random Field model was used (Silverman and Cooper, 1988).

Range images, which are input sources for 3D recognition systems, have been segmented using clustering procedures. Each pixel in a range image only points to a location in 3D space so it is easier to cluster than normal images. On the other hand, the addition of complex features that do not represent distance would require data manipulation for successful results (Jain and Flynn, 1993).

Bezdek (2013) pioneered the use of fuzzy logic in clustering data, which is beneficial when there are no clear boundaries. He explained that pattern recognition is made up of the data, the data search method and the structure. Structure here refers to the way the data is organized so that the links and relationships between the values can be recognized. The Expectation-Maximization (EM) algorithm is a foundational algorithm in clustering that provides a way to compute estimates on incomplete or hidden data. By treating the cluster assignment of regions as unknown data, this iterative algorithm can estimate the probability that the region belongs to a certain cluster. With each iteration, it incrementally improves the segmentation process (Dempster *et al*., 1977)

The segmentation of positron emission tomography (PET) images can be difficult due to the sensitivity of relevant algorithms to a number of parameters. This algorithm is important in medical applications for the detection of tumours and lesions. The goal of this paper was to present an improved statistical approach other than simple thresholding methods or manual representation techniques. Current segmentation methods required prior data and had poor detection of small lesions and this new method addressed those limitations. The proposed approach combined Expectation Maximization Gaussian mixture model (EMGMM) for tissue classification along with a Markov random field (MRF) and a Gibbs distribution for segmentation smoothening. The results of the investigation displayed a more robust and accurate detection of tumours than traditional fixed thresholding, particularly in cases involving differing signal-to-background ratios (SBRs) and small lesions (Layer *et al*., 2015).

The point distribution model (PDM) is a statistical model that was developed by representing objects as a set of points and assessing how the positions of those points change over multiple training examples (Cootes *et al*., 1992). Active shape models are able to repeatedly adjust a shape model to find outlines of objects in an image. This method uses PDMs to identify those shapes in an image. During the fitting process, an optimization technique estimates initial parameters for scale, shape, position and orientation of objects in an image to improve the models. Those points are updated iteratively to better fit the correct boundaries of those objects (Cootes and Taylor, 1992).

Constraints are applied to the model's parameters to ensure that the generated shapes are consistent with the training examples provided. Active shape models are capable of modelling resistors and human body parts including the hands and the heart. Those landmark points must be accurately placed accurately on the training images to ensure that the model can depict those points without the noise from point location errors. This approach has practical applications in object detection, automated image analysis and medical interpretation. At the time, existing model-based vision methods were too flexible and would produce inaccurate matches. This led to the development of active shape models that would only change shape in realistic ways for that object class (Cootes *et al*., 1995).

With the increased complexity of images together with the vast differences between the objects in those images, the obstacles associated with the extraction of features and accurate segmentation became unsurmountable. Before deep learning, random forest and semantic text on forests were commonly used to develop segmentation classifiers. The introduction of convolutional neural networks (CNNs), which were able to process any image

size, laid the foundation for future advancements in deep learning segmentation techniques (Yu *et al.*, 2023).
It is important to note that although modern deep learning methods are more advanced and efficient, it cannot be ignored that the classical methods served as a foundation for the later introduced methods. The limitations of these classical methods ultimately led to the move to deep learning techniques.

METHODOLOGY

This study combined three MCDM tools into a three-phase framework for evaluating different deep learning architectures (Figure 1). The first phase of the framework was designed using a DEMATEL method, while the second phase of the framework was designed using a FAHP method. The last phase of the framework was designed using a fuzzy VIKOR method. The details of the phases are presented as follows:

Phase I: Enhanced DEMATEL with temporal dynamics

This phase is used to map the causal interdependencies among criteria. It accounts for temporal influence propagation and feedback loops. The phase consists of five steps: multi-round direct influence assessment, dynamic normalization with influence decay, total relation matrix with feedback amplification, enhanced prominence-relation analysis, and dynamic impact-relation mapping. The multi-round direct influence assessment step is used to implement the DEMATEL-inspired iterative refinement process (Figure 1).

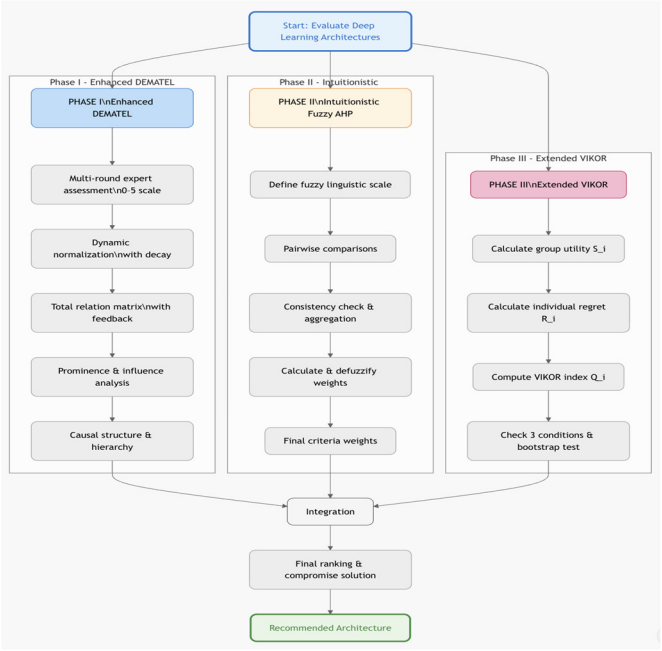


Figure 1: Proposed framework

i.Each expert provides an initial influence matrix using an enhanced scale, as shown in (Table 1).

ii.Calculate inter-rater reliability (Kendall's W).
iii. If W < 0.5, conduct facilitated discussion and a second round.

iv. Aggregate the results using weighted geometric mean.

The dynamic normalisation, with influence decay, step is used to compute the temporal attenuation of influence criteria. Equation 1 gives the expression for normalisation with influence decay (Table 1).

Table 1: Enhanced scale for criteria evaluation

0	No influence
1	Negligible influence (threshold effect)
2	Low influence (minor cascading)
3	Moderate influence (noticeable impact)
4	High influence (significant driver)
5	Critical influence (systemic determinant)

$$N_{dyn} = A.s.e^{-\lambda t} \quad (1)$$

Where *s* denotes the normalisation scalar, λ denotes a decay parameter, and *t* denotes the temporal lag between criteria activation.

The total relation matrix, with feedback amplification, step is used to incorporate reinforcing loops into the enhanced DEMATEL method. Equation (2) gives the expression of the reinforcing loop.

$$T = N(I - \beta N)^{-1} \quad (2)$$

Where β denotes the feedback damping coefficient. A low β value denotes a stronger feedback suppression.

The enhanced prominence-relation analysis step is used to generate extended metrics for the decision-making problem. The extended metrics consist of prominence (Equation 3), net influence (Equation 4), centrality score (Equation 5), and influence leverage (Equation 6).

$$P_i = R_i + C_i \quad \text{(total system engagement)} \quad (3)$$

$$NI_i = R_i - C_i \quad \text{(causal vs. effect role)} \quad (4)$$

$$CS_i = \sqrt{R_i^2 + C_i^2} \quad \text{(Geometric importance)} \quad (5)$$

$$IL_i = \frac{R_i}{C_i + \epsilon}$$ (Diver potency ratio) (6)

The last step of in this phase, i.e., dynamic impact-relation mapping, is used to map the quadrant-based strategic classification for criteria.

Quadrant I (high P, high NI): Core drivers - require intensive optimization.

Quadrant II (high P, low NI): Effect amplifiers - monitor closely.

Quadrant III (low P, low NI): Independent factors - evaluate separately.

Quadrant IV (low P, high NI): Latent drivers - potential optimization targets.

Output: Causal structure map, influence hierarchy, temporal dependency chains.

The DEMATEL procedure is presented as follows:

1.Round 1: Full direct-influence DEMATEL matrices (0–5 scale) + linguistic pairwise comparisons for intuitionistic fuzzy AHP + self-assessed confidence score (0–1) for each matrix.

2.Response rate.

3.Feedback report: Anonymous summary statistics (mean, median, standard deviation) for every cell, highlighting cells with SD >1.0 or Kendall's W <0.7 at the matrix level.

4.Round 2: Experts received the anonymized feedback report and were asked to revise only where they wished; they again provided a confidence score for the revised matrix.

The inclusion of fuzzy sets in the DEMATEL method address the limitation that is inherent in conventional DEMATEL method implementation. This study implemented two rounds of evaluation to reduce the workload of experts.

Phase II: Intuitionistic fuzzy AHP with dynamic consensus

This phase is used to determine criteria weights. The current study used intuitionistic fuzzy sets (IFS) to capture hesitation and non-membership values of criteria. Table 2 shows the IFS for the current decision-making problem. The implementation of this phase consists of seven steps: the first step deals with stating the intuitionistic fuzzy linguistic scale for the decision-making problem. This study used Triangular Intuitionistic Fuzzy Numbers (TIFN) to capture vagueness and uncertainty in experts' judgments (Equation 7). Table 2 shows the TIFN for the current decision-making problem. The hesitation degree for each judgment is expressed as Equation (8).

Table 2: IFS for the decision-making problem.

Linguistic term	TFN	μ	v
Equal importance	(1, 1, 1)	0.9	0.05
Weak importance	(1, 2, 3)	0.75	0.15
Moderate importance	(2, 3, 4)	0.7	0.2
Strong importance	(3, 4, 5)	0.65	0.25
Very strong	(4, 5, 6)	0.6	0.3
Absolute Importance	(5, 6, 7)	0.55	0.35

$$\tilde{A} = [(l, m, u);\ \mu, v]$$ (7)

$$\pi = 1 - \mu - v$$ (uncertainty) (8)

Where (l, m, u) denotes triangular fuzzy number, μ denotes membership degree (confidence) and v denotes non-membership degree (rejection).

The second step under this phase is used to generate information about the intuitionistic fuzzy pairwise comparison of the decision-making problem (Equation 9). This matrix consists of information about criteria preference strength (fuzzy), confidence level (membership), rejection level (non-membership) and hesitation (residual uncertainty).

$$\tilde{A}^{IF} = \left[\tilde{a}_{ij}^{IF}\right]$$ (9)

Consistency checking with hesitation adjustment represents the third steps under this phase. Equation (10) gives the hesitation-aware consistency ratio.

$$S(\tilde{a}_{ij}) = \frac{l+2m+u}{4}.(\mu - v)$$ (10)

The fourth step under this phase deals with consensus-based aggregation. Equation (11) gives the weighted aggregation by expertise and agreement.

$$\tilde{w}_{ij}^{agg} = \frac{\sum_{k=1}^{K} \alpha_k.\theta_k.\tilde{w}_{ij}^{k}}{\sum_{k=1}^{K} \alpha_k.\theta_k}$$ (11)

Where α_k and $\tilde{w}_{ij}^{k}$ denote expert *k*'s domain weight and agreement coefficient (based on similarly to group median), respectively.

The derivation of the intuitionistic fuzzy weight for criteria represents the fifth step under this phase. This study used geometric mean method to compute the weight (Equations 12 and 13).

$$\tilde{r}_i^{IF} = \left[\left(\prod_{j=1}^{n} l_{ij}\right)^2, \left(\prod_{j=1}^{n} m_{ij}\right)^2, \left(\prod_{j=1}^{n} u_{ij}\right)^2\right]; \mu_i, v_i \quad (12)$$

$$\tilde{w}_i^{IF} = \tilde{r}_i^{IF} \otimes \left(\sum_{i=1}^{n} \tilde{r}_i^{IF}\right)^{-1} \quad (13)$$

The sixth step deals with the defuzzification of the intuitionistic fuzzy weight with uncertainty preservation (Equation 14).

$$W_i^{crisp} = \frac{l_i + 2m_i + u_i}{4} . (\mu_i - v_i) . \sqrt{1 - \pi_i} \quad (14)$$

$\sqrt{1-\pi_i}$ penalises high hesitation judgements.

The last step deals with adaptive DEMATEL-fuzzy AHP integration. Equation (15) gives the performance-based dynamic weighting.

$$w_i^{Final} = \alpha(t) . w_i^{IF-AHP} + [1 - \alpha(t)] . \frac{CS_i}{\sum_{i=1}^{n} CS_i} \quad (15)$$

$$\alpha(t) = 0.5 + 0.3 . cos\left(\frac{\pi . t}{T_{eval}}\right) \quad (16)$$

Phase III: Extended VIKOR with Regret-aversion optimisation

This study used an extended VIKOR method to generate compromise ranks for the architectures. Extension is made to standard VIKOR method using the concept of prospect theory principles and multi-stakeholder perspectives. Equation (17) gives the group utility with an aspiration gap.

$$S_i = \sum_{j=1}^{n} w_j . \frac{f_j^{asp} - f_{ij}}{f_j^{asp} - f_j^{tol}} \quad (17)$$

Where: f_j^* (ideal) denotes the best possible value, f_j^{asp} (aspiration) denotes the realistically achievable target, f_j^{tol} (tolerable) denotes the minimum acceptable threshold, and f_j^- (anti-ideal) denotes the worst observed value

Equation (18) shows the individual regret for the architectures with prospect adjustment for alternative.

$$R_i = \left[w_j . \left(\frac{f_j^{asp} - f_{ij}}{f_j^{asp} - f_j^{tol}}\right)^{\gamma}\right] \quad (18)$$

where γ denotes the loss-aversion parameter (lower values = higher regret-aversion).

Equation (19) gives the VIKOR index with stakeholder preferences using multi-v strategy for diverse perspectives. This study considered three variants of v: 0.3 (maximising worst-case performance),0.5 (standard VIKOR), and 0.7 (maximising average performance).

$$Q_i^v = v . \frac{S_i - S^*}{S^- - S^*} + (1 - v) \frac{R_i - R^*}{R^- - R^*} \quad (19)$$

Equation (20) gives the expression used to optimise the VIKOR outputs.

$$Q^v\left(A^{(2)}\right) - Q^v\left(A^{(1)}\right) \geq DQ(1 + \sigma Q) \quad (20)$$

The VIKOR results optimisation is based on three conditions: acceptable advantage (enhanced), acceptance stability (extended) and strategic alignment.

Condition 1: Acceptable advantage (enhanced)

Where σQ is the coefficient of variation in Q-scores, requiring larger gaps when scores are volatile.

Condition 2: Acceptance stability (extended)

$A^{(1)}$ must rank top 2 of 3 measures (S, R, $Q^{(0.5)}$)

Condition 3: Strategic alignment

$A^{(1)}$ m must not be in the worst quartile for any criterion with $w_j > 0.10$

The three conditions above are used to generate a compromise solution set based on a bootstrapped ranking stability concept. This concept is explained as followed:

i. Perform 1000 bootstrap iterations, resampling expert judgments.
ii. Calculate rank frequency distribution for each alternative.
iii. Compute ranking confidence: $RC_i = P(rank_i \leq 3)$.

Table 3: Evaluation criteria.

Criterion	Description	Orientation
Accuracy (C1)	Measures the overall proportion of correct predictions made by the model in tasks like classification or segmentation.	Higher-the -better
Sensitivity (C2)	Evaluates the model's ability to correctly identify positive cases, crucial for minimising missed diagnoses.	Higher-the -better
Specificity (C3)	Assesses the model's performance in correctly identifying negative cases, helping to reduce false positives in diagnostic applications.	Higher-the -better
Precision (C4)	Quantifies the proportion of positive identifications that are actually correct, important for ensuring reliability in medical predictions.	Higher-the -better
F1-Score (C5)	Provides a harmonic mean of precision and recall, balancing the trade-off between false positives and false negatives in imbalanced medical datasets.	Higher-the -better
Area under the ROC curve (C6)	Measures the model's ability to distinguish between classes across various threshold settings, commonly used for binary classification in imaging.	Higher-the -better
Dice coefficient (C7)	Evaluates overlap between predicted and ground-truth segmentations, particularly relevant for tasks like organ or lesion segmentation.	Lower-the -better
Inference time (C8)	Represents the computational time required for the model to process and output results on new images, essential for real-time clinical applications.	Lower-the -better
Model size (C9)	Indicates the resource footprint of the architecture (number of parameters or memory usage), affecting deployment on hardware-constrained medical devices.	lower-the -better
Training time (C10)	Measures the duration needed to train the model, impacting feasibility in iterative development or resource-limited environments.	Higher-the -better
Robustness to variations (C11)	Assesses how well the model performs under noise, artifacts, or variations in imaging modalities, ensuring reliability in diverse clinical settings.	Higher-the -better
Interpretability (C12)	Evaluates the model's ability to provide explainable outputs, which is vital for clinician trust and regulatory approval in medical imaging.	Higher-the -better
Ethical and fairness considerations (C13)	Assesses whether the model minimizes bias and adheres to ethical guidelines in medical decision-making.	Higher-the -better

The following decision rules are used to analyse the compromise solution set.

If all conditions are met and $RC_i > 0.80$: Recommend as sole solution.

If Condition 3 is violated: Exclude $A^{(1)}$, evaluate $A^{(2)}$.

If Conditions 1 and 2 are met, but $RC_i < 0.80$: Recommend compromise set $\{A^{(1)}, A^{(2)}\}$.

Otherwise: Recommend top-k set where $Q^v(A^{(k)}) - Q^v(A^{(1)}) < 1.5DQ$

RESULTS AND DISCUSSION

This study evaluates six deep learning architectures for medical image segmentation: Swin-Unet, nnU-Net, Attention U-Net, TransUNet, U-Net, and UNETR. These architectures are constructed from convolutional models. The advantages of these architectures in handling complex medical imaging tasks varies from one criterion to another. The current study used the proposed framework to evaluate these architectures based on criteria in Table 3. Assessment of the criteria was carried out using 12 expert judgments. Five of the experts have experience in diagnostic radiology, while four of the experts have experience in medical physics with focus on deep learning. The remaining three experts are clinicians. Each expert has at least 10 years' experience of professional practice and a minimum of second degree. They were selected based on their experience in medical image segmentation and willingness to participate in two rounds and provide self-assessed confidence scores. During data collection, this study ensured anonymity.

For example, responses were collected via a secure online platform. Table 4 presents the aggregated direct influence matrix for the 13 evaluation criteria. This study obtained the information using a weighted geometric mean based on expert confidence scores. Table 5 shows the normalized influence matrix with temporal dynamics applied. Equation (1) was used to compute temporal attenuation of criteria influences. Table 6 contains the total relation matrix incorporating feedback amplification for causal relationships among criteria (see Equation 2). Table 7 shows the extended metrics results. This study generated the values in this table using Equations (3) and (4). From the results in (Table 7), this study observed that five criteria represent the core drivers for evaluating the deep learning architectures for medical image segmentation – C3, C5, C10, C11, and C12, while C1 and C7 are considered as effect amplifiers.

Table 4: Aggregated matrix (enhanced weighted aggregation).

Criteria	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13
C1	-	0.7434	2.2478	3.481	1.0058	1.2187	2.2566	1.4956	2.2478	2.7289	3	2.5248	1.7434
C2	2.2478	-	0.2332	2	2.5335	1.4956	2.4956	1.5044	0.2332	1.5102	2.2478	1.0087	1.7376
C3	1.2391	2.758	-	3.0292	2.0087	3.2711	1.7434	2.2711	3.2624	2.2332	2.758	3.2332	0.9854
C4	2.2566	0.4898	2.5248	-	1.5335	0.2566	3.0233	1.7726	0.5102	2.2536	1.2624	1	0.4665
C5	1.758	1.4956	1.9942	2	-	1.2332	3.5044	3	2.5335	1.4869	1.519	2.481	2.0292
C6	2.2478	2.0233	2.0058	1.2478	1.2711	-	1.519	1.2332	2.0146	2.4898	1.7289	2.2711	2.4956
C7	3.0146	2.4898	3	1.481	2.7668	1.4752	-	1.2332	3	1.0292	1.2624	1.4752	0.7755
C8	1	0.7726	3.5335	1.2187	0.5102	0.9767	3.2711	-	0.7376	3.2711	3	3.2566	1
C9	2.7434	2	0.9854	3.2187	1.2711	2	3	1.9854	-	2.7522	0.7434	2.0087	0.4956
C10	2.5248	3	1.9767	0.9913	2.5044	1.7464	3	2.4898	0.723	-	1.2624	2.5044	2.0233
C11	2.7609	1.5102	1	1.519	2.5248	2.4898	1.7668	3	3	0.758	-	3.2711	2.4752
C12	2.7376	0.5102	2.9913	2.9708	3.2332	1.2566	0.9942	2.7522	0.723	3.2478	3.2332	-	3.277
C13	3.000	0.7522	1.2566	2.5131	3.7376	3.000	1.9854	1.2624	2.7434	0.758	1.7464	0.2566	-

Table 5: Normalized matrix (decay parameter of 0.1).

Criteria	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13
C1	-	0.0191	0.0706	0.1094	0.0316	0.0347	0.0709	0.0425	0.0706	0.0858	0.0853	0.0718	0.0548
C2	0.0639	-	0.0066	0.0515	0.0652	0.0385	0.0642	0.0387	0.0073	0.0389	0.0578	0.0260	0.0447
C3	0.0352	0.0867	-	0.0779	0.0517	0.1028	0.0496	0.0714	0.0928	0.0575	0.0710	0.1016	0.031
C4	0.0642	0.0139	0.0793	-	0.0482	0.0073	0.095	0.0557	0.016	0.0641	0.0359	0.0284	0.0147
C5	0.05	0.0425	0.0627	0.0515	-	0.0351	0.1101	0.0772	0.0796	0.0423	0.0391	0.0705	0.0522
C6	0.0578	0.0636	0.063	0.0355	0.0399	-	0.0391	0.0388	0.0518	0.0641	0.0445	0.0646	0.071
C7	0.0857	0.0641	0.0853	0.0421	0.0869	0.0419	-	0.0317	0.0772	0.0265	0.0325	0.038	0.02
C8	0.0284	0.0199	0.0909	0.0383	0.0145	0.0278	0.093	-	0.019	0.0842	0.0943	0.0838	0.0314
C9	0.078	0.0515	0.031	0.1011	0.0327	0.0569	0.0853	0.0565	-	0.0865	0.0234	0.0631	0.0141
C10	0.065	0.0943	0.0621	0.0282	0.0644	0.0449	0.0853	0.0708	0.0186	-	0.0325	0.0712	0.0636
C11	0.0868	0.0429	0.0284	0.0432	0.065	0.0641	0.0555	0.0853	0.0943	0.0238	-	0.1028	0.0704
C12	0.086	0.0145	0.0851	0.0764	0.0919	0.0323	0.0283	0.0783	0.0206	0.0836	0.0832	-	0.0843
C13	0.0772	0.0194	0.0323	0.079	0.0962	0.0853	0.0565	0.0325	0.0706	0.0216	0.0497	0.0081	-

Table 6: Total relation matrix (β = 0.9).

Criteria	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	R
C1	0.107	0.093	0.165	0.202	0.126	0.111	0.180	0.136	0.152	0.174	0.168	0.168	0.126	1.906
C2	0.132	0.048	0.074	0.115	0.126	0.088	0.138	0.100	0.067	0.098	0.115	0.092	0.094	1.287
C3	0.150	0.161	0.106	0.181	0.151	0.179	0.169	0.169	0.176	0.158	0.164	0.204	0.114	2.082
C4	0.135	0.070	0.148	0.071	0.113	0.064	0.171	0.120	0.080	0.127	0.100	0.102	0.068	1.370
C5	0.148	0.111	0.154	0.145	0.089	0.108	0.210	0.160	0.157	0.130	0.122	0.161	0.119	1.814
C6	0.145	0.123	0.141	0.121	0.120	0.068	0.133	0.117	0.122	0.140	0.119	0.146	0.132	1.624
C7	0.169	0.124	0.162	0.129	0.159	0.107	0.097	0.111	0.148	0.107	0.108	0.124	0.083	1.629
C8	0.120	0.086	0.171	0.121	0.099	0.096	0.181	0.082	0.095	0.158	0.166	0.168	0.097	1.638
C9	0.168	0.115	0.122	0.183	0.115	0.119	0.182	0.136	0.072	0.167	0.103	0.148	0.081	1.712
C10	0.159	0.157	0.150	0.120	0.150	0.115	0.184	0.152	0.098	0.085	0.117	0.159	0.131	1.776
C11	0.189	0.112	0.129	0.146	0.156	0.138	0.167	0.174	0.175	0.122	0.092	0.197	0.144	1.941
C12	0.187	0.090	0.182	0.175	0.182	0.113	0.146	0.172	0.112	0.174	0.172	0.106	0.158	1.968
C13	0.161	0.080	0.114	0.159	0.166	0.144	0.151	0.109	0.141	0.100	0.119	0.093	0.061	1.597
C	1.970	1.369	1.817	1.869	1.751	1.449	2.107	1.738	1.595	1.740	1.664	1.868	1.408	-

The independent criteria are C2, C4, and C8. The remaining criteria are latent drivers – i.e., C6, C9, and C13.

Phase II: Enhanced intuitionistic fuzzy AHP

Table 8 shows the derivation of criteria weights using intuitionistic fuzzy AHP. This study used TIFN to capture membership, non-membership, and hesitation degrees in pairwise comparisons (Equations 7 and 8). Equations 10 to 14 were used to generate the criteria weights. The results in this table showed that C1 to C3 are the principal criteria for evaluating the deep learning architectures for medical image segmentation. Table 9 shows criteria weights after integrating intuitionistic fuzzy AHP with

Table 7: Comprehensive metrics for the criteria.

Criteria	R	C	P	NI	CI	IL	Quadrant
C1	1.91	1.97	3.88	-0.064	2.741	0.968	Effect amplifier
C2	1.29	1.37	2.66	-0.081	1.879	0.941	Independent
C3	2.08	1.82	3.90	0.264	2.764	1.145	Core driver
C4	1.30	1.87	3.17	-0.499	2.317	0.733	Independent
C5	1.81	1.75	3.56	0.062	2.521	1.036	Core driver
C6	1.62	1.45	3.07	0.175	2.177	1.121	Latent driver
C7	1.63	2.11	3.74	-0.478	2.663	0.773	Effect amplifier
C8	1.64	1.74	3.38	-0.100	2.388	0.942	Independent
C9	1.71	1.60	3.31	0.117	2.340	1.074	Latent driver
C10	1.78	1.74	3.52	0.036	2.486	1.021	Core driver
C11	1.94	1.66	3.60	0.277	2.557	1.167	Core driver
C12	1.97	1.87	3.84	0.101	2.714	1.054	Core driver
C13	1.60	1.41	3.01	0.189	2.130	1.134	Latent driver

Table 8: Detailed intuitionistic fuzzy weight derivation.

Criteria	L	M	U	μ	v	Defuzzified
C1	0.046	0.122	0.259	0.567	0.167	0.050
C2	0.076	0.155	0.304	0.567	0.167	0.063
C3	0.088	0.163	0.308	0.548	0.206	0.056
C4	0.033	0.082	0.175	0.546	0.167	0.032
C5	0.047	0.094	0.192	0.447	0.212	0.023
C6	0.050	0.093	0.185	0.381	0.280	0.010
C7	0.024	0.055	0.119	0.368	0.224	0.008
C8	0.029	0.057	0.122	0.306	0.307	0.000
C9	0.028	0.053	0.111	0.264	0.380	-0.006
C10	0.018	0.036	0.080	0.248	0.356	-0.004
C11	0.018	0.035	0.077	0.209	0.445	-0.009
C12	0.016	0.030	0.067	0.184	0.515	-0.011
C13	0.013	0.024	0.054	0.167	0.567	-0.011

Table 9: Final weighted criteria with uncertainty analysis.

Criteria	Final weight	Lower bound	Upper bound	Uncertainty range
C1	0.1747	0.1363	0.2250	0.0887
C2	0.1729	0.1312	0.2265	0.0953
C3	0.1535	0.1165	0.2012	0.0847
C4	0.1050	0.0774	0.1398	0.0624
C5	0.0902	0.0614	0.1251	0.0636
C6	0.0626	0.0383	0.0912	0.0529
C7	0.0619	0.0423	0.0857	0.0435
C8	0.0450	0.0285	0.0646	0.0361
C9	0.0378	0.0236	0.0547	0.031
C10	0.0322	0.0214	0.0451	0.0237
C11	0.0283	0.0192	0.0395	0.0203
C12	0.0233	0.0168	0.0314	0.0146
C13	0.0126	0.0095	0.0164	0.0069

DEMATEL (see Equations 15 and 16). An α value of 0.80 was used to implement the adaptive DEMATEL-fuzzy AHP model (Equation 14). The results obtained confirm sensitivity (C1), specificity (C2), and accuracy (C3) as top-weighted criteria for evaluating the deep learning architectures for medical image segmentation.

Phase III: Enhanced VIKOR ranking

Table 10 shows the performance scores of the six deep learning architectures for the thirteen evaluation criteria. This information are pre-processed data obtained from the experts. The information in this table served as input for the adaptive VIKOR implementation. The multi-level utility thresholds for the adaptive VIKOR method implementation are presented in (Table 11). The values in this table are based on the consensus among the 12 experts. This study used Equation (17) to compute group utility with aspiration gap and Equation (18) individual regret with prospect adjustment. Table 12 shows the ranking confidence of the compromise solutions based on bootstrap iterations. Bootstrap iteration was used to resampling expert

Table 10: Performance Matrix.

Model	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13
U-Net	0.924	0.877	0.943	0.921	0.908	0.926	0.885	0.700	1.000	1.000	0.844	0.800	0.822
Attention U-Net	0.934	0.895	0.945	0.912	0.913	0.936	0.893	0.700	1.000	1.000	0.859	0.845	0.836
nnU-Net	0.952	0.935	0.972	0.939	0.938	0.971	0.908	0.700	1.000	1.000	0.900	0.740	0.872
TransUNet	0.924	0.911	0.946	0.914	0.903	0.930	0.877	0.700	1.000	1.000	0.846	0.823	0.825
UNETR	0.897	0.884	0.930	0.908	0.879	0.927	0.867	0.700	1.000	1.000	0.835	0.781	0.805
Swin-Unet	0.959	0.921	0.971	0.925	0.918	0.955	0.921	0.700	1.000	1.000	0.886	0.832	0.854

Table 11: Thresholds for the criteria.

Criteria	Ideal	Asp	Tol	Anti
Benefit	1	0.9	0.8	0.7
Non-benefit	0.1	0.2	0.4	0.5

Table 12: Bootstrap ranking confidence.

Model	Ranking Confidence (P (rank ≤ 3))
Swin-Unet	1
nnU-Net	1
Attention U-Net	0.934
TransUNet	0.066
U-Net	0
UNETR	0

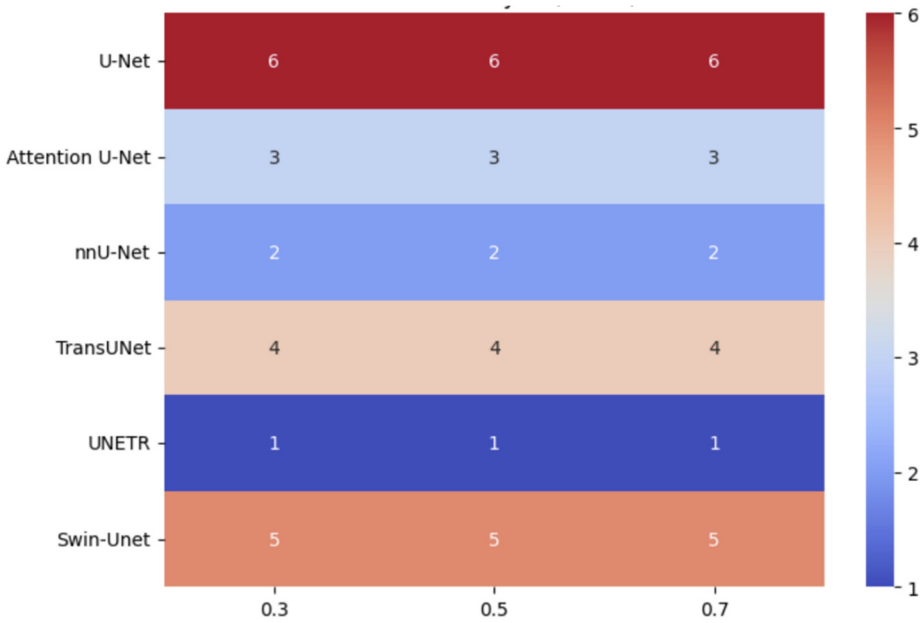


Figure 2: Sensitivity analysis (ranks)

judgments to assess ranking stability (Shekhovtsov and Sałabun, 2020). The following results were obtained: Swin-Unet: Q = 0.000, Confidence = 1.000 (optimal compromise solution); nnU-Net: Q = 0.0070, Confidence = 1.000 (best compromise solution); Attention U-Net. The high confidence for the top three architectures validates the robustness of the proposed framework. This study checked the VIKOR conditions and observed the following:
i.Condition 1 (Advantage): Failed (0.0070 < 0.3311).
ii.Condition 2 (Stability): Passed (top in 2/3 measures: S, R, Q);
iii.Condition 3 (Alignment): Passed (not in worst quartile for high-weight criteria).

Based on these observations, this study recommended the following top-3 compromise set: Swin-Unet, nnU-Net, Attention U-Net for medical image segmentation. The sensitivity results in (Figure 2) showed that the VIKOR

ranked the architectures as UNETR >> nnU-Net >> Attention U-Net >> TransUNet >> Swin-Unet >> classic U-Net.

Conclusions

This study presents an adaptive multi-stage multi-criteria decision-making framework that addresses the selection of optimal deep learning architectures for medical image segmentation. The framework combined DEMATEL, intuitionistic fuzzy AHP, and adaptive VIKOR. Sensitivity, specificity and accuracy were identified as top-weighted criteria for evaluating the deep learning architectures for medical image segmentation. The VIKOR method results generated top-3 compromise set: Swin-Unet, nnU-Net, Attention U-Net for medical image segmentation. Based on the sensitivity analysis report, UNETR and classic U-Net were the most and least suitable deep learning architecture, respectively, for medical image segmentation. Future research directions include extending the model to include real-time performance monitoring, studying adaptive threshold adjustment mechanisms through deployment feedback, and looking at multi-modal fusion strategies for evaluating hybrid architectures across a wide range of medical imaging domains and practice settings. Further study will be carried out to investigate the impacts of the framework's parametric settings on the selection of the deep learning architecture for medical image segmentation.

REFERENCES

Abhisheka, B., Biswas, S. K., Purkayastha, B., Das, D., & Escargueil, A. (2024). Recent trend in medical imaging modalities and their applications in disease diagnosis: A review. *Multimedia Tools and Applications, 83*(14), 43035–43070.

Adams, R., & Bischof, L. (1994). Seeded region growing. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 16*(6), 641–647.

Angelis, A. N. (2017). *Multiple criteria decision analysis for assessing the value of new medical technologies: Researching, developing and applying a new value framework for the purpose of health technology assessment* [Doctoral dissertation, London School of Economics and Political Science].

Bezdek, J. C. (2013). *Pattern recognition with fuzzy objective function algorithms*. Springer Science & Business Media.

Canny, J. (1986). A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-8*(6), 679–698.

Cootes, T. F., & Taylor, C. J. (1992). Active shape models - 'smart snakes.' In *Proceedings of the British Machine Vision Conference 1992* (pp. 28.1–28.10). Springer-Verlag London Limited.

Cootes, T. F., Taylor, C. J., Cooper, D. H., & Graham, J. (1992). Training models of shape from sets of examples. In *Proceedings of the British Machine Vision Conference 1992* (pp. 2.1–2.10). Springer-Verlag London Limited.

Cootes, T. F., Taylor, C. J., Cooper, D. H., & Graham, J. (1995). Active shape models—Their training and application. *Computer Vision and Image Understanding, 61*(1), 38–59.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological), 39*(1), 1–38.

Ghribi, F., & Hamdaoui, F. (2025). Innovative deep learning architectures for medical image diagnosis: A comprehensive review of convolutional, recurrent, and transformer models. *The Visual Computer, 41*(13), 11603–11628.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-means clustering algorithm. *Applied Statistics, 28*(1), 100.

Jain, A. K., & Flynn, P. J. (1993). *Three-dimensional object recognition systems*. Elsevier Science Inc.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys, 31*(3), 264–323.

Kass, M., Witkin, A., & Terzopoulos, D. (1988). Snakes: Active contour models. *International Journal of Computer Vision, 1*(4), 321–331.

Kittler, J., & Illingworth, J. (1985). On threshold selection using clustering criteria. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-15*(5), 652–655.

Kulak, O., Goren, H. G., & Supciller, A. A. (2015). A new multi criteria decision making approach for medical imaging systems considering risk factors. *Applied Soft Computing, 35*, 931–941.

Layer, T., Bijos, A., Lorenz, E., Kuwert, T., & Pöppl, S. J. (2015). PET image segmentation using a Gaussian mixture model and Markov

Li, M., Jiang, Y., Zhang, Y., & Zhu, H. (2023). Medical image analysis using deep learning algorithms. *Frontiers in Public Health, 11*, Article 1273253.

Ojha, J., Presacan, O., Lind, P. G., Monteiro, E., & Yazidi, A. (2025). Navigating uncertainty: A user-perspective survey of trustworthiness of AI in healthcare. *ACM Transactions on Computing for Healthcare, 6*(3), 1–32.

Osher, S., & Sethian, J. A. (1988). Fronts propagating with curvature-dependent speed: Algorithms based on Hamilton-Jacobi formulations. *Journal of Computational Physics, 79*(1), 12–49.

Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics, 9*(1), 62–66.

Ridler, T. W., & Calvard, S. (1978). Picture thresholding using an iterative selection method. *IEEE Transactions on Systems, Man, and Cybernetics, 8*(8), 630–632.

Schachter, B. J., Davis, L. S., & Rosenfeld, A. (1979). Some experiments in image segmentation by clustering of local feature values. *Pattern Recognition, 11*(1), 19–28.

Shekhovtsov, A., & Sałabun, W. (2020). A comparative case study of the VIKOR and TOPSIS rankings similarity. *Procedia Computer Science, 176*, 3730–3740.

Silverman, J. F., & Cooper, D. B. (1988). Bayesian clustering for unsupervised estimation of surface and texture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 10*(4), 482–495.

Suzuki, K. (2017). Overview of deep learning in medical imaging. *Radiological Physics and Technology, 10*(3), 257–273.

Vincent, L., & Soille, P. (1991). Watersheds in digital spaces: An efficient algorithm based on immersion simulations. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 13*(6), 583–598.

Weeks, A. R., Myler, H. R., & Apley, L. F. (1994, May). Adaptive local thresholding algorithm that maximizes the contour features within the thresholded image. In E. R. Dougherty, J. Astola, & H. G. Longbotham (Eds.), *IS&T/SPIE 1994 International Symposium on Electronic Imaging: Science and Technology* (p. 230). SPIE.

Weiner, E. B., Dankwa-Mullan, I., Nelson, W. A., & Hassanpour, S. (2025). Ethical challenges and evolving strategies in the integration of artificial intelligence into clinical practice. *PLOS Digital Health, 4*(4), Article e0000810.

Wong, A. K. C., & Sahoo, P. K. (1989). A gray-level threshold selection method based on maximum entropy principle. *IEEE Transactions on Systems, Man, and Cybernetics, 19*(4), 866–871.

Yu, Y., Wang, C., Fu, Q., Kou, R., Huang, F., Yang, B., Yang, T., & Gao, M. (2023). Techniques and challenges of image segmentation: A review. *Electronics, 12*(5), Article 1199.