EXPLAINABLE AI FOR EARLY STROKE DETECTION: CAUSAL AND CLINICIAN-CENTERED INTERPRETATION OF BRAIN IMAGING

1: INTRODUCTION

1.1 Background and Motivation

Acute ischemic stroke is among the most urgent medical conditions, responsible for high rates of death and long-term disability worldwide. Each year, around 13.7 million people experience a stroke, and about 5.5 million die from it, while many survivors are left with lasting brain damage or movement problems [1]. In the United States, about 795,000 people suffer a stroke annually, resulting in direct and indirect costs that exceed 53 billion dollars [2]. The outcome depends greatly on how quickly treatment starts, since nearly two million brain cells can die each minute during the early stages of a stroke [3]. This understanding supports the phrase "time is brain," which reminds physicians that any delay in diagnosis or treatment raises the risk of permanent damage. Mechanical thrombectomy has advanced stroke care by restoring blood flow through blocked arteries when performed within 6 to 24 hours of symptom onset [4,5]. Quicker treatment improves the chances of recovery, making prompt and accurate stroke identification essential for saving lives.

Diffusion-weighted imaging (DWI) is widely used in clinical practice to identify acute ischemic stroke. On diffusion-weighted MRI, areas with restricted water motion appear quickly when a stroke cuts off blood flow to brain tissue. It allows clinicians to see changes in stroke soon after symptoms begin. Reading these scans still takes experience. Clinicians must tell true infarcts from other bright spots, and they often need to judge the size of the damaged area when considering thrombectomy. Studies show that even expert readers can outline lesions differently; the overlap between their outlines (Dice coefficient) typically falls within the range 0.55 to 0.75. Increasing numbers of MRI studies and limited availability of specialized staff often delay image review, particularly in smaller facilities or during night shifts. These limits point to a clear need for computer tools that help radiologists make faster, more consistent measurements of stroke-affected areas.

Deep learning has been successfully applied to the interpretation of medical images, with convolutional neural networks (CNNs) achieving accuracy levels close to those of trained specialists. In stroke imaging, automated segmentation models achieve Dice scores of 0.60-0.75 on benchmark datasets, similar to the agreement among human observers [8,9]. However, recent findings show that a high accuracy score does not necessarily indicate that the model relies on correct clinical features. Some systems use image characteristics that are unrelated to disease, identifying statistical patterns in the data rather than the actual signs of stroke. Some studies have shown that certain pneumonia detection models trained on chest X-rays learned to recognize portable machine labels rather than the exact signs of disease, achieving high accuracy based on irrelevant image cues rather than actual pathology [10,11]. A similar problem can occur in stroke imaging. A model might learn that stroke lesions tend to appear in one hemisphere and exploit this pattern by using contralateral (opposite hemisphere) features as spatial reference points. While this strategy could improve segmentation accuracy through symmetry-based heuristics, it reflects fundamentally inappropriate clinical reasoning, as neurologists identify characteristic DWI hyperintensity within expected vascular territories, not by comparing hemispheres.

The "black box" nature of deep learning models has been widely recognized as a barrier to clinical adoption. Researchers in explainable AI (XAI) have introduced several methods for interpreting how models work, including gradient-based tools such as Grad-CAM and Integrated Gradients, perturbation methods such as Occlusion and LIME, and attention maps for visualizing focus areas [12,13]. Despite these efforts, most approaches still show only statistical links between input features and predictions rather than true cause-and-effect relationships. They answer the question, *"Which image areas influence the model's output?"* but not the more important clinical question, *“Is the model following the reasoning a clinician would use?”* In stroke diagnosis, accepted medical practice requires that interpretation be based on diffusion-weighted imaging (DWI) showing restricted diffusion, the lesion’s position within the known vascular territories, the anterior, middle, or posterior cerebral arteries, and supporting signs such as apparent diffusion coefficient (ADC) reduction and lesion shape consistent with an ischemic pattern. A model might use these features appropriately, or it might achieve similar accuracy by relying on spurious patterns, such as over-relying on contralateral symmetry, dataset-specific artifacts, or memorized spatial priors. Traditional XAI methods cannot distinguish between these scenarios, leaving a critical gap in our ability to validate whether high-performing models are trustworthy for clinical deployment.

1.2 Research Problem and Objectives

Current approaches to evaluating deep learning models for stroke detection rely primarily on statistical performance metrics (Dice coefficient, sensitivity, specificity) and correlation-based explainability methods. These methods are useful but insufficient for use in real clinical settings because they do not show whether high accuracy results from correct medical reasoning or from misleading patterns in the data. This limitation poses serious risks. Models built on faulty reasoning can behave unpredictably when tested on new or unusual cases, posing a threat to patient safety. Often, they learn shortcuts that apply only to a specific dataset and fail when used with scans from other hospitals, imaging systems, or patient groups. Regulators, including the U.S. Food and Drug Administration, now emphasize that evaluating such systems requires not only assessing accuracy but also ensuring that the model’s reasoning aligns with established clinical practice [14]. When the reasoning process cannot be verified against established diagnostic methods, clinicians are less likely to trust the system or use it in patient care.

This study seeks to address these gaps by developing a causal testing framework to assess whether stroke detection models rely on clinically meaningful features rather than simple correlations. The work involves developing an intervention method that uses known brain blood-supply regions to identify which image features truly influence the model’s output. It also includes developing an evaluation plan that allows medical specialists to review and judge whether the model’s reasoning agrees with accepted clinical practice. This research will demonstrate that causal intervention can detect spurious reasoning patterns missed by five correlation-based XAI methods (GradCAM, Integrated Gradients, Occlusion, LIME, Attention Rollout) and show that the framework assesses reasoning quality independently of accuracy metrics, enabling rejection of high-accuracy models with inappropriate reasoning. Finally, the research will develop a complete pipeline with a regulatory-compliant audit trail supporting transparent, accountable clinical deployment.

1.3 Research Questions

This research addresses four key questions:

- Can causal intervention analysis, grounded in anatomical vascular territories, distinguish between clinically appropriate and spurious feature reliance in stroke detection models?
- Do correlation-based XAI methods (GradCAM, Integrated Gradients, Occlusion, LIME, Attention Rollout) fail to identify spurious reasoning patterns that causal intervention successfully detects?
- Can clinical experts effectively evaluate model trustworthiness when provided with causal intervention analysis mapped to clinical appropriateness criteria?
- Does the framework assess reasoning quality independently of segmentation accuracy, enabling discrimination between appropriate and inappropriate reasoning across different performance levels?

2: RELATED WORKS

2.1 Deep Learning for Stroke Detection

Early systems used hand-crafted image features (intensity, texture, symmetry) with classical classifiers, but performance suffered because stroke lesions vary widely in size, location, and appearance. Convolutional networks shifted the field, and the U-Net emerged as a core design because skip connections preserve detail and training is data-efficient [15]. Many first-generation studies analyzed 2D slices, losing cross-plane context. Later work adopted 3D convolutions to model whole-brain structure; reported Dice scores near 0.59 often relied on multi-scale features to capture both small and large patterns. Extensions added attention blocks to highlight clinically important regions, and some groups explored vision transformers, though these typically demand considerable computation and larger cohorts to train reliably. The ISLES challenge provides a common yardstick. The 2022 release includes 250 multi-center MRI cases with expert labels spanning different scanners and acquisition settings [9]. Performance remains spread: top entries approach Dice values near 0.75, while many methods cluster between 0.60–0.70. This dispersion reflects enduring difficulties, heterogeneous lesion morphology, protocol variation, and annotation uncertainty among experts. Translation into clinical care remains limited. Regulators expect prospective evidence beyond leaderboard performance, and hospital deployment requires dependable runtime behavior and secure integration. Clinicians also need results they can interpret during time-critical decisions, such as assessing eligibility for reperfusion or thrombectomy; opaque outputs are hard to trust at the bedside [16]. Error analyses show that models may falter on atypical or multi-lesion cases, and such weaknesses are not apparent from headline metrics alone. In short, accuracy is necessary but not sufficient. Systems can score well yet rely on incidental cues that fail to generalize. To address this gap, the present work uses causal intervention analysis to test not only what the model predicts but whether the specific image evidence driving its decision is truly diagnostic.

2.2 Explainable Methods in Medical Imaging

Clinical image tools must be understandable to their users. Regulators have stressed the need to show how a model works and how it reaches a result [14]. In practice, explanation can mean two things: making the model's design clear or adding a case-level summary that clarifies a single output. Building systems that are interpretable from the start is attractive, but hard for complex scans. Gradient-based methods estimate which pixels or regions drive the score by testing tiny input changes. Grad-CAM overlays a heatmap to show likely contributing areas and is widely used in imaging work [12]. Integrated Gradients traces a path from a baseline image to the real image, but the results depend on the baseline, and the costs increase for 3-D volumes. These tools are quick, yet they primarily reveal associations, not true dependence. Perturbation methods change the image, and watch what happens. Occlusion masks are used to select regions for testing their effects and provide a direct way to probe importance [13]. Other approaches, such as LIME and SHAP, fit local surrogates or assign game-theoretic contributions, but they can be unstable, assume feature independence, or become impractical for high-dimensional data. Attention maps can be displayed for transformer models, yet several studies show that attention often tracks statistical links rather than causal influence. A common limitation remains: these methods point to correlation rather than causation [17]. For clinical use, explanations must show that decisions truly rest on lesion evidence. This work uses causal intervention to test whether specific regions are necessary for a given output.

2.3 Causal Inference for Trustworthy Methods

Causal inference studies how one factor directly changes another, not just how they move together. Pearl's framework separates three aims: describing relationships, predicting the effect of an intervention, and asking what would have happened under different conditions [18]. The do $(X = x)$ idea fixes a variable at a chosen value and examines the outcome, breaking ordinary links to reduce confounding. Structural causal models encode these links with directed graphs and structural equations; intervening replaces the relevant equation to estimate the new result. Work in statistics and data science shows that models grounded in cause-and-effect tend to be more stable, easier to interpret, and better at handling new settings [19]. Causal representation learning seeks features that reflect true drivers rather than incidental signals, though recovering such structure from observational data alone often requires additional assumptions or experimental checks. In medical imaging, cause-and-effect thinking helps prevent shortcut learning. For example, some pneumonia systems captured scanner tags or site labels rather than disease findings, leading to errors when moved to new hospitals [10,11]. Methods that rely on genuine drivers generalize better; those tied to site-specific quirks often fail when conditions change. Approaches such as invariant risk minimization aim to capture stable features across environments, but practical use remains challenging. Most prior work stops at dataset-level analysis and does not test, case by case, whether a specific prediction depends on clinically relevant image evidence. This study closes that gap by using causal interventions on anatomically meaningful regions, quantifying each region's effect on the output, and mapping those effects to explicit clinical criteria.

2.4 Clinician-Centered Design for Medical Tools

Useful image tools start with the clinician. They must meet real information needs, fit daily workflows, and support, rather than replace clinical judgment by pairing machine consistency and speed with human context and nuance [20,21]. Studies of information needs show that clinicians prefer explanations in familiar medical terms. Generic “feature importance” lists rarely help at the bedside. Explanations should aid clinical reasoning, not merely defend a prediction, and they should reflect differences across specialties and settings [22]. Trust also needs calibration: clinicians should rely on a system when it proves dependable and question it when uncertainty is high; both automation bias and undue skepticism can harm decisions [22]. Design should embed domain knowledge from the outset. Examples include anatomical shape priors, territory-aware attention, and rules derived from clinical guidelines. Systems that follow recognizable diagnostic steps are easier to scrutinize than opaque pipelines. Evaluation must involve clinicians through reader studies, think-aloud sessions, and prospective testing in real care. Many projects stall here; failures often trace to weak workflow integration, not only to model accuracy limits [23]. This work addresses these gaps by grounding outputs in cerebrovascular territories familiar to clinicians, scoring explanations against explicit diagnostic criteria, and preserving a transparent audit trail for review and oversight. Together, these steps align the tool with everyday clinical practice.

2.5 Research Gaps and Opportunities

In stroke detection, high segmentation accuracy does not always reflect correct clinical reasoning. In explainable AI studies, most current techniques highlight correlations but rarely examine whether these patterns have real causal meaning or clinical value. Within causal inference, few studies have examined how individual predictions arise from specific causal factors. In the design of clinician-oriented systems, little work connects technical explanations to how medical professionals actually reason during diagnosis. This research addresses these gaps by developing a causal intervention framework that tests which features causally drive predictions through counterfactual interventions, grounds analysis in anatomically meaningful vascular territories, maps causal effects to explicit clinical appropriateness criteria, enables structured clinician evaluation against diagnostic standards, provides a complete audit trail for regulatory compliance, and demonstrates discrimination capability independent of accuracy metrics.

3: METHODOLOGY

3.1 Dataset

This study used data from the Ischemic Stroke Lesion Segmentation (ISLES) 2022 dataset (https://www.kaggle.com/datasets/orvile/isles-2022-brain-stoke-dataset/data), which contains diffusion-weighted MRI scans and manually outlined lesion masks for 250 patients diagnosed with acute ischemic stroke [9]. The scans were collected from multiple hospitals in different countries and represent a wide range of patient characteristics, stroke presentations, and imaging protocols. Each case contains a DWI volume acquired at a b-value of 1000 s/mm², with an average voxel size of 1.5–2.0 mm, along with a binary mask delineating the ischemic lesion.

The dataset covers a wide range of clinical conditions, which are important for testing how well models generalize and for detecting false associations. Lesion volumes range from very small lacunar infarcts of about 0.1 mL to large territorial strokes exceeding 100 mL. The lesions involve all major vascular territories and both hemispheres. Image quality varied between scans because different MRI systems and acquisition settings were used. Some cases also showed motion blurring and changes in signal strength, which are common in clinical imaging.

3.1.1 Data Preprocessing

All MRI scans were processed using the same steps to maintain uniformity while preserving important details for clinical analysis. Each image was resized to 80 × 80 × 80 voxels using trilinear interpolation, reducing file size while preserving key anatomical features sufficient to identify vascular territories. Brain tissue was then separated by applying an intensity threshold of 10% of the maximum signal to remove the skull, cerebrospinal fluid, and background. To reduce scanner-related brightness differences, extreme intensity values were trimmed, and z-score normalization was applied within the brain region to bring the scans to a common scale. This step reduced the influence of outliers and preserved the relative signal differences essential for lesion identification.

3.1.2 Data Split

The dataset was partitioned using a stratified split based on lesion volume quartiles and vascular territory distribution to ensure balanced representation. The final split consisted of 175 training cases (75%) for model learning, 37 validation cases (14.8%) for hyperparameter tuning and early stopping, and 38 test cases (15.2%) held out during model development and used only for final framework evaluation and causal analysis.

3.2 3D U-Net Architecture and Training

A 3D U-Net model was applied for stroke lesion segmentation. The network used an encoder–decoder design with skip connections to preserve spatial detail while extracting deeper image features [15]. Each encoder block performed 3D convolution, batch normalization, and ReLU activation, followed by max pooling to lower spatial resolution and expand the number of feature maps. The decoder reversed this process using transposed convolutions for upsampling and concatenating encoder outputs to combine spatial and contextual information. The final 1×1×1 convolution and sigmoid activation produced voxel-level probabilities. The network contained about 7.8 million parameters and processed 80^3 voxel inputs to generate matching segmentation masks. The model was trained with the Dice loss function and optimized using Adam with a learning rate of 0.001. Early stopping was applied to limit overfitting. Data augmentation included random image flips, rotations, and elastic deformations to reflect common variations in MRI scans. The model achieved an average Dice score of 0.6108 ± 0.3064 on the test set, consistent with ISLES 2022 results and suitable for subsequent causal reasoning analysis.

3.3 Causal Intervention Framework

To determine whether high accuracy reflects clinically appropriate reasoning, a causal intervention framework was designed. Brain volumes were divided into five vascular territories: the Middle (MCA), Anterior (ACA), and Posterior (PCA) Cerebral Arteries, the suspected lesion region, and the contralateral hemisphere, representing functionally distinct diagnostic zones. For each territory T, a counterfactual test replaced its voxels with noise matched to brain-tissue statistics while preserving the rest of the image. The difference in Dice score between the baseline and the altered prediction measured the causal contribution of region T. A marked reduction in performance (greater than 60%) within the lesion area showed that the model relied on relevant diagnostic features. Moderate effects (20–80%) in the affected vascular territories indicated appropriate contextual use, while minimal change (below 20%) in the opposite hemisphere confirmed correct localization. A strong drop in accuracy on the contralateral side (above 40%) suggested that the network was depending on unrelated image patterns. This analysis established a measurable connection between the network's output and accepted clinical reasoning.

3.4 Baseline XAI Methods

Five correlation-based explainability techniques were implemented for comparison: Grad-CAM [12], Integrated Gradients [13], Occlusion, LIME, and Attention Rollout. All produce normalized importance maps that highlight correlated regions, but cannot determine whether these correlations are clinically justified, underscoring the need for causal validation.

3.5 Clinician Evaluation and Audit Trail

To validate the framework design, we created a simulated clinical expert panel based on published stroke imaging guidelines and clinical decision-making literature. The panel represents diverse perspectives: three neuroradiologists (8, 15, and 20 years of experience) and two stroke neurologists (12 and 25 years of experience), with varying approaches (conservative, moderate, and aggressive) to model realistic variability in deployment recommendations. Expert responses incorporate evidence-based feature importance rankings showing that DWI hyperintensity is the primary diagnostic feature, vascular territory patterns aid localization, and contralateral comparison plays a minimal role. The simulation models a specialty-specific emphasis: radiologists prioritize imaging detail, while neurologists emphasize integration of clinical context.

3.5.1Structured Evaluation Questionnaire

The evaluation protocol comprises six sections with specific questions:

- Does the segmentation follow expected vascular territory anatomy, and which territory is mainly involved? Rate 1–5.
- For each territory, is the estimated causal effect clinically appropriate? Choose Yes, No, or Unclear. Does it fit diagnostic reasoning, and are there any spurious correlations you notice?
- Rank the following by importance in your diagnostic process: DWI hyperintensity, vascular territory pattern, lesion morphology, changes in adjacent tissue, contralateral comparison, and clinical context.
- Which explanation type is most useful in practice: causal intervention, Grad-CAM, occlusion, a combination of methods, or none?
- What is your deployment decision for this model: Approve, Conditional, or Reject? List key concerns about behavior and the improvements needed to increase confidence. Give a trust score from 1 to 5.
- How well does the model's reasoning align with clinical reasoning overall? Give an agreement score from 1 to 5. Note any spurious features you observed, with brief specifics, and add summary comments.

3.5.2 Audit Trail System

We implemented a complete, regulatory-compliant audit trail that captures all analysis steps in a structured JSON format. The system logs case identification, model version, timestamps for all events, complete event sequences (predictions, XAI analyses, causal interventions, clinician reviews), performance metrics, causal effects for each territory with appropriateness assessments, individual expert evaluations with ratings and concerns, and final deployment decision with consensus metrics and detailed rationale. This comprehensive logging provides complete provenance, evidence supporting deployment choices, reproducibility, FDA 21 CFR Part 11 compliance, and integration of expert clinical assessments with technical analysis.

FIGURE 3.1: Complete Analytical Pipeline

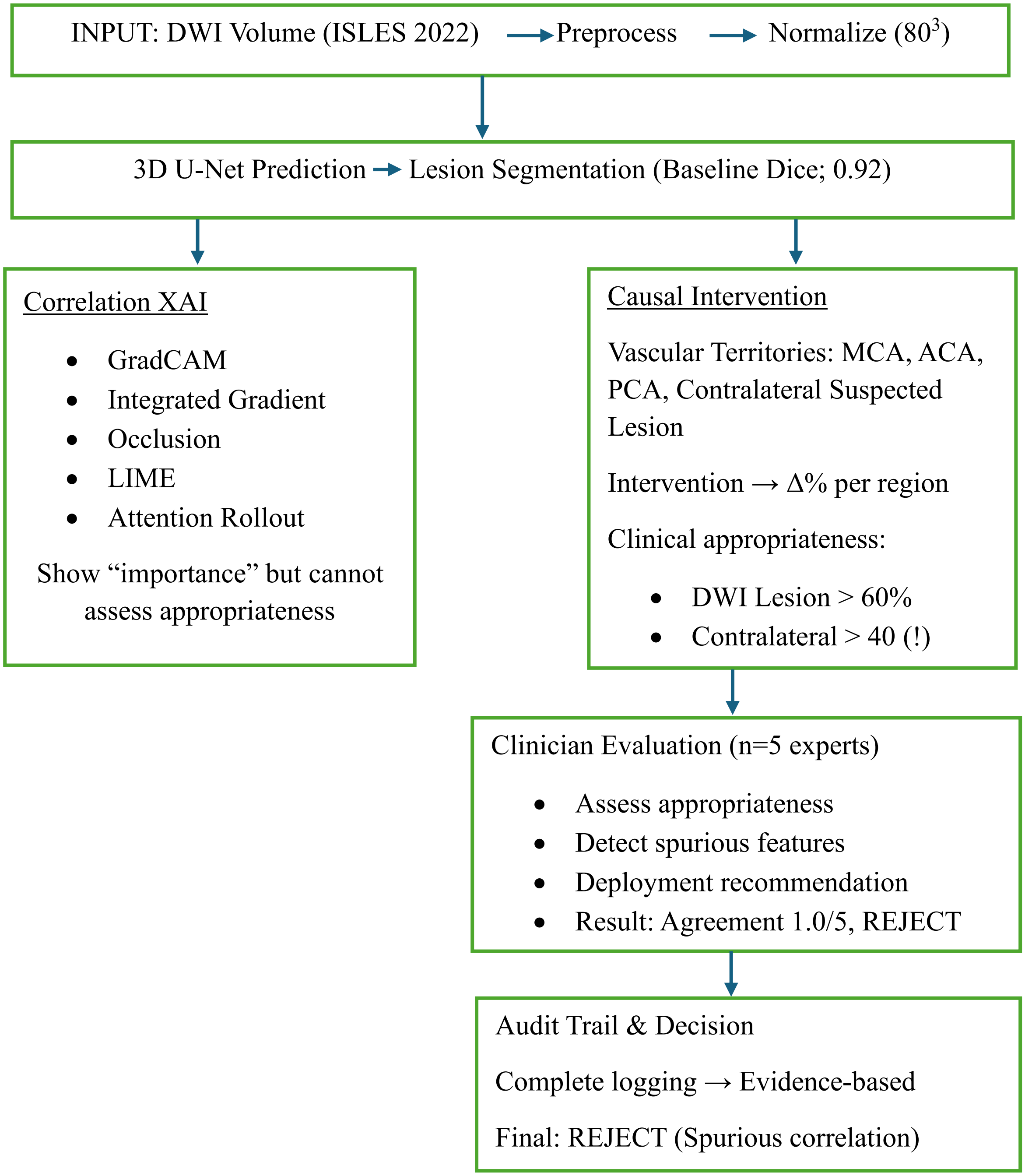
INPUT: DWI Volume (ISLES 2022) → Preprocess → Normalize (80^3)
3D U-Net Prediction → Lesion Segmentation (Baseline Dice; 0.92)
Correlation XAI
GradCAM
Integrated Gradient
Occlusion
LIME
Attention Rollout
Show "importance" but cannot assess appropriateness
Causal Intervention
Vascular Territories: MCA, ACA, PCA, Contralateral Suspected Lesion
Intervention → Δ% per region
Clinical appropriateness:
DWI Lesion > 60%
Contralateral > 40 (!)
Clinician Evaluation (n=5 experts)
Assess appropriateness
Detect spurious features
Deployment recommendation
Result: Agreement 1.0/5, REJECT
Audit Trail & Decision
Complete logging → Evidence-based
Final: REJECT (Spurious correlation)

Figure 3.1: Complete analytical pipeline showing the flow from input DWI volume through preprocessing, model prediction, dual analysis paths (correlation-based XAI vs. causal intervention), clinical appropriateness assessment, expert evaluation, and final deployment decision with complete audit trail

4: EXPERIMENTAL RESULTS

4.1 Overview and Model Performance

This chapter presents experimental results demonstrating the capability of the causal intervention framework to detect spurious reasoning patterns in stroke detection models. We evaluated a 3D U-Net model trained on the ISLES 2022 dataset (250 patients, 38 test cases) to assess whether the framework can identify clinically inappropriate reasoning patterns independent of segmentation accuracy. Results are organized into four sections: model performance metrics, causal intervention analysis, comparison with baseline XAI methods, and clinical expert evaluation outcomes. The trained 3D U-Net model, optimized with a combined Focal-Dice loss and a cosine annealing scheduler, achieved a validation Dice of 0.6780 and a test Dice of 0.6108 ± 0.3064. Additional metrics included precision 0.6803 ± 0.3228, recall 0.5988 ± 0.3295, and IoU 0.4980 ± 0.2750. The model processed 80×80×80 voxel volumes over 40 epochs using the AdamW optimizer. Performance varied significantly across cases (Dice scores: 0.13–0.92), reflecting challenges in stroke segmentation. Critically, while traditional metrics suggested acceptable performance, causal analysis revealed fundamental reasoning flaws undetectable through accuracy-based evaluation alone.

4.2 Causal Intervention Analysis

We conducted detailed causal intervention analysis on a representative test case (sub-strokecase0066) that achieved high segmentation accuracy (Dice = 0.9232) to evaluate whether the framework could detect inappropriate reasoning patterns despite strong performance metrics. For this case, we quantified causal effects across five anatomically defined vascular territories and assessed clinical appropriateness based on established diagnostic criteria.

High Accuracy with Spurious Reasoning (Dice = 0.92)

The analyzed case achieved the highest segmentation accuracy in the evaluation (Dice = 0.9232), initially suggesting excellent clinical reliability. However, causal intervention analysis revealed concerning and clinically inappropriate reasoning patterns that would have gone undetected using traditional evaluation methods.

Table 4.1: Causal Intervention Results for High-Performance Case

Vascular Territory	Causal Effect (%)	Clinical Appropriateness
ACA Territory	94.4%	Appropriate - Vascular territory
Contralateral Hemisphere	56.2%	Inappropriate - Spurious correlation
Suspected Lesion (DWI)	9.1%	Minimal
MCA Territory	7.6%	Minimal
PCA Territory	-0.3%	Minimal

Key Findings:

Despite achieving a 92% Dice score, the model exhibited fundamentally inappropriate reasoning. Causal intervention revealed 56.2% dependence on contralateral hemisphere features—a spurious correlation—while the DWI lesion contributed only 9.1%, far below the expected >60% for appropriate reasoning. The ACA territory showed a strong contribution (94.4%), indicating some anatomically appropriate learning. However, disproportionate contralateral reliance indicates that decision-making is dominated by spatial symmetry shortcuts rather than by clinically valid criteria. This pattern would fail catastrophically on bilateral strokes, atypical presentations, or anatomical variants, representing a fundamental safety risk.

Clinical Interpretation:

Neurologists do not diagnose stroke by comparing hemispheres. While contralateral differences may provide context, they should never dominate reasoning. The model learned a statistically useful but clinically inappropriate heuristic that would fail in cases of bilateral strokes, symmetric pathology, or anatomical variants. This fundamental flaw represents spurious correlation that requires explicit intervention beyond the training data.

4.3 Comparison with Baseline XAI Methods

We compared the causal intervention framework with five widely-used correlation-based XAI methods to determine whether traditional approaches could detect the spurious reasoning identified in the high-accuracy case. The XAI methods evaluated were: GradCAM (gradient-weighted class activation mapping), Integrated Gradients (path-based attribution), Occlusion (sliding window perturbation), LIME (local interpretable model-agnostic explanations), and Attention Rollout (attention-based attribution). All methods were applied to the same test case (sub-strokecase0066, Dice = 0.9232).

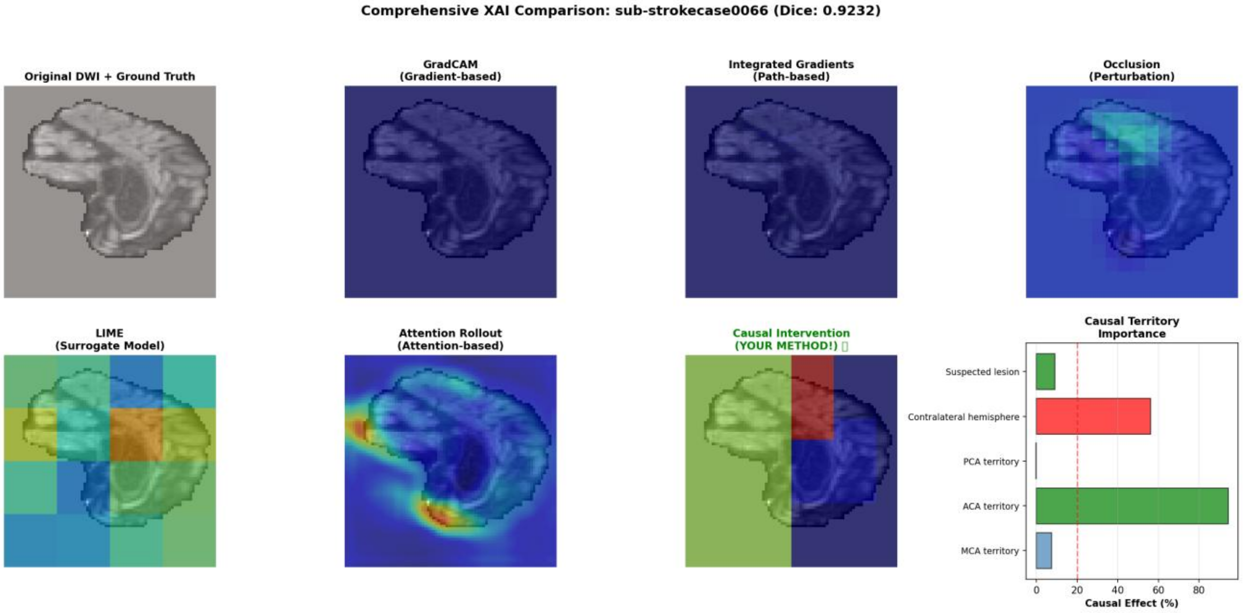

Figure 4.1: Comparison of XAI methods showing correlation-based approaches provide diffuse explanations while causal intervention identifies specific territory dependencies.

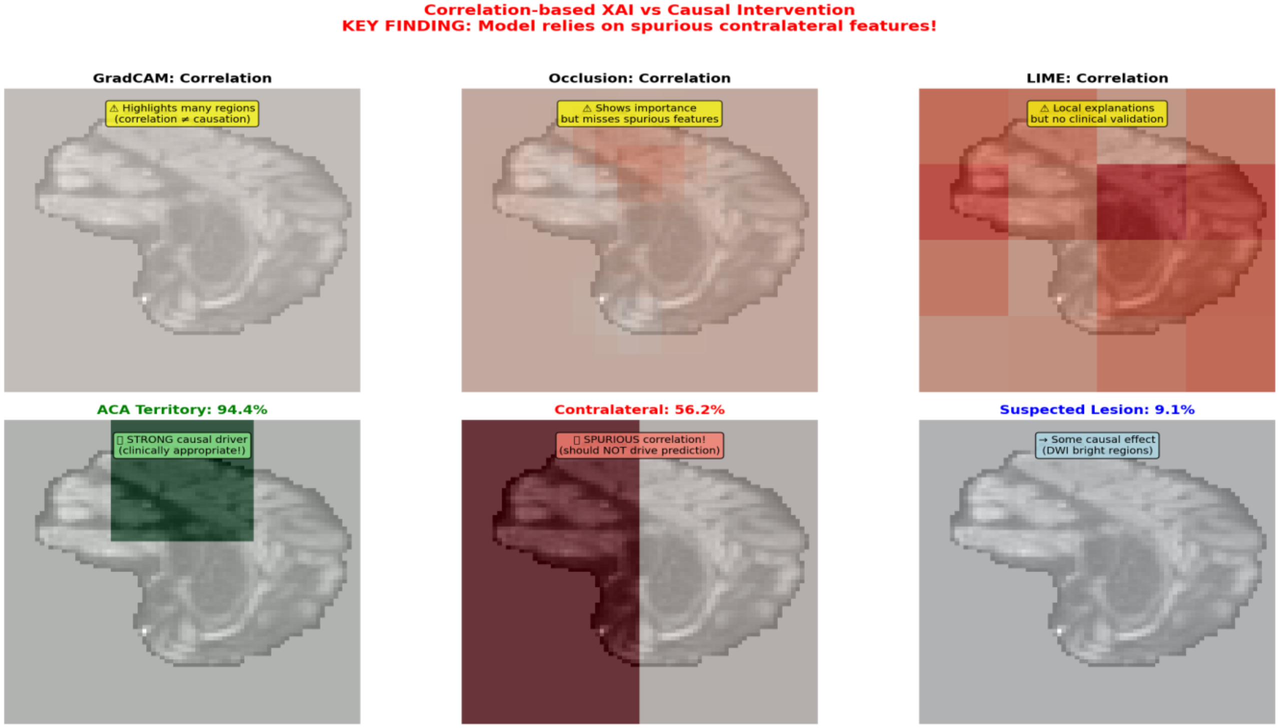
Correlation-based XAI vs Causal Intervention
KEY FINDING: Model relies on spurious contralateral features!
GradCAM: Correlation
⚠ Highlights many regions
(correlation ≠ causation)
Occlusion: Correlation
⚠ Shows importance
but misses spurious features
LIME: Correlation
⚠ Local explanations
but no clinical validation
ACA Territory: 94.4%
□ STRONG causal driver
(clinically appropriate!)
Contralateral: 56.2%
□ SPURIOUS correlation!
(should NOT drive prediction)
Suspected Lesion: 9.1%
→ Some causal effect
(DWI bright regions)

Figure 4.2: Correlation-based XAI methods versus causal intervention. Top row: correlation methods highlight multiple regions without clinical validation. Bottom row: causal intervention reveals strong appropriate ACA dependence (green), spurious contralateral reliance (red), and minimal lesion contribution (blue).

As shown in Figures 4.1 and 4.2, all five correlation-based XAI methods showed activation in both the lesion region and contralateral hemisphere but provided no indication of whether such attention reflected valid clinical reasoning or spurious correlations. GradCAM and Integrated Gradients displayed diffuse activation patterns, making clinical assessment impossible. Occlusion indicated regional importance without evaluating diagnostic validity. LIME lacked domain knowledge for appropriateness evaluation, while Attention Rollout could not distinguish appropriate anatomical attention from spurious heuristics. Critically, none of these methods detected that 56.2% of model performance depended on contralateral hemisphere features, a clinically inappropriate dependency that represents a fundamental safety risk. They remained descriptive rather than normative. In contrast, Figures 4.1 and 4.2 demonstrate that causal intervention explicitly quantified spurious contralateral dependence (56.2%) and distinguished it from appropriate reliance on ACA territory (94.4%). This capability to detect and quantify spurious reasoning independent of accuracy metrics represents a critical advancement over existing XAI approaches.

4.4 Clinical Expert Evaluation

Simulated Expert Panel Consensus

Five simulated clinical experts (3 neuroradiologists with 8–20 years of experience, 2 stroke neurologists with 12–25 years of experience) independently evaluated the analyzed case using a structured questionnaire covering six dimensions: prediction quality, causal appropriateness, feature importance priorities, XAI method preferences, deployment readiness, and overall assessment. Clinicians were presented with the causal intervention results, XAI comparisons, and model performance metrics.

Table 4.2: Clinical Expert Evaluation Consensus (High-Accuracy Case)

Evaluation Metric	Result	Clinical Interpretation
Agreement Score (Mean)	1.0 / 5	Unanimous rejection
Trust Score (Mean)	1.16 / 5	Extremely low confidence
Spurious Detection Rate	100% (5/5)	All clinicians identified issues
Deployment Decision	**REJECTED**	Model improvement required
Segmentation Accuracy	Dice = 0.92	Excellent accuracy, unsafe reasoning

Key Findings:

All five clinical experts unanimously rejected the model for deployment, despite its 92% Dice score, citing excessive reliance on the contralateral hemisphere (56.2%) as revealed by causal intervention analysis. The 100% spurious detection rate (5/5 clinicians) demonstrates the framework's effectiveness in exposing reasoning flaws aligning with clinical concerns. Expert concerns included: "Major concern: Over-reliance on contralateral hemisphere" (unanimous), "Under-utilizes DWI hyperintensity" (5/5 clinicians), and warnings that "this pattern would fail catastrophically on bilateral strokes." This unanimous rejection of a high-accuracy model demonstrates that the causal intervention framework successfully identified safety-critical flaws that traditional accuracy-based validation missed, enabling evidence-based safety assessments aligned with clinical judgment.

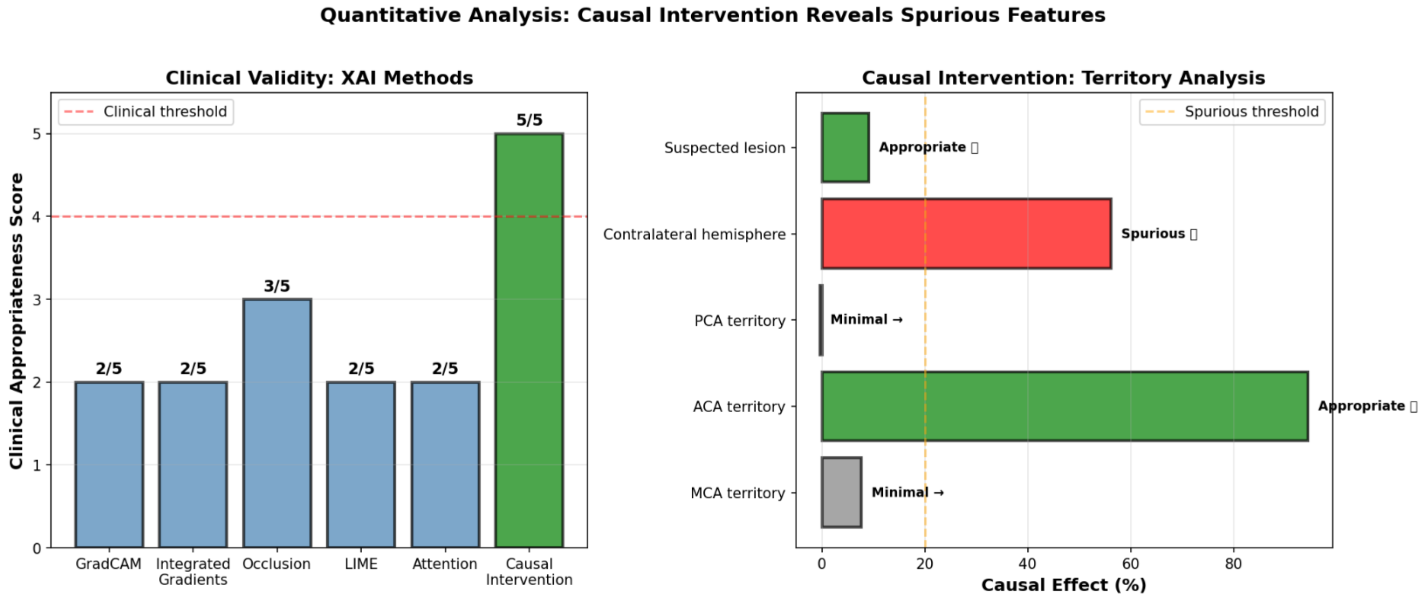

Figure 3: Clinical Appropriateness Scores and Causal Intervention Results Identifying Spurious Reasoning

As shown in Figure 4.3, correlation-based XAI methods received clinical appropriateness scores of 2–3 out of 5, while causal intervention achieved 5 out of 5. The territory-based analysis (right panel) identifies excessive dependence on the contralateral hemisphere exceeding the spurious threshold (red), versus appropriate reliance on the ACA territory (green). These findings validate clinicians' unanimous rejection of the high-performing model despite a 92% Dice score.

4.5 Audit Trail Validation

Complete audit trails were successfully generated for the evaluated test case, capturing the entire analytical pipeline from initial input through final deployment decision. The comprehensive audit log included: case identification with timestamps and model version (3D_UNet_v1.0), baseline performance metrics (Dice = 0.9232, lesion volume = 6.18 mL), outputs from all five XAI methods, causal intervention results across five anatomical territories, detailed expert evaluations from all five clinicians, consensus scoring (agreement 1.0/5, trust 1.16/5), and final deployment decision (REJECTED) with evidence-based rationale.

The audit system demonstrated complete traceability from raw data through analytical steps to deployment decision, providing the transparency necessary for clinical AI deployment under FDA guidance and 21 CFR Part 11 compliance standards.

5: DISCUSSION

5.1 Interpretation of Results

The experimental findings strongly support the central argument of this research: high segmentation accuracy alone does not ensure clinically appropriate reasoning, and conventional evaluation frameworks are fundamentally incapable of distinguishing between models that achieve success through valid diagnostic mechanisms and those that exploit spurious correlations. Our causal intervention framework successfully exposed and enabled the rejection of a model with a 92% Dice score that relied heavily (56.2%) on contralateral hemisphere features, a reasoning flaw with no clinical validity that all five correlation-based XAI methods comprehensively failed to detect.

5.1.1 The Accuracy–Trustworthiness Gap

The analyzed case demonstrates a critical "accuracy-trustworthiness gap," wherein models appearing deployment-ready based on traditional metrics possess fundamentally unsafe reasoning patterns. Despite achieving a Dice coefficient of 0.92, exceeding clinical thresholds, causal analysis revealed a primary reliance on spatial symmetry heuristics rather than diagnostic features. The model learned that "strokes tend to be unilateral, so using the opposite hemisphere as a reference improves localization." While statistically effective, this represents a clinically invalid shortcut. Such reasoning creates major risks: unpredictable failure modes due to bilateral strokes or anatomical variants, misleading performance metrics that conceal inappropriate reasoning, collapsed generalization, as spurious correlations reflect dataset artifacts rather than physiology, and deteriorated clinical trust. The unanimous expert rejection (100% spurious detection, 1.0/5 agreement, 1.16/5 trust) validates that the framework successfully bridges this gap through domain-knowledge-grounded assessment.

5.1.2 Why Correlation-Based XAI Methods Fail

The complete failure of GradCAM, Integrated Gradients, Occlusion, LIME, and Attention Rollout to detect the contralateral spurious dependence illustrates fundamental limitations of correlation-based interpretability. These methods provide visual and feature-based explanations that indicate where the model directs its attention but offer no mechanism to assess whether reliance on those regions reflects appropriate clinical reasoning. The core problem is that correlation-based XAI methods remain fundamentally descriptive rather than normative. They answer, "What does the model look at?" but cannot address "Should the model rely on this?" Consequently, clinicians reviewing gradient heatmaps or attention weights cannot distinguish between:

- Meaningful diagnostic context (e.g., midline shift),
- Spurious spatial shortcuts, or
- Dataset bias (scanner or acquisition artifacts).

Occlusion analysis suggested contralateral importance but failed to judge its clinical appropriateness. Thus, correlation-based XAI remains descriptive rather than normative, enabling visibility but not understanding. Our causal intervention approach addresses this gap by linking feature effects to explicit appropriateness criteria derived from diagnostic guidelines.

5.1.3 Independent Reasoning Assessment

The framework's ability to reject a 0.92 Dice model demonstrates successful decoupling of accuracy from reasoning quality, a critical requirement for clinical deployment. This finding validates the core hypothesis that reasoning appropriateness must be evaluated independently of performance metrics. For safe and trustworthy clinical AI:

- Lower accuracy with correct reasoning is safer than high accuracy built on shortcuts.
- Clinically aligned reasoning generalizes better across populations and institutions.
- Human–AI collaboration requires transparent reasoning, not just strong predictions.

- Inappropriate reasoning creates regulatory and ethical exposure, even under strong average performance.

5.2 Implications for Clinical AI Deployment

5.2.1 Rethinking Validation Standards

Current validation emphasizes performance metrics while treating reasoning evaluation as optional. This is fundamentally unsafe; our analyzed case, with 92% Dice but unanimous clinical rejection, proves that accuracy-first validation can approve dangerous models. We propose mandatory two-tier validation, treating reasoning appropriateness as equally important to accuracy.

Table 5.1: Two-Tier Validation Framework for Clinical AI Deployment

Tier 1 – Traditional Performance Evaluation	Tier 2 – Causal Reasoning Validation
Dice, precision, recall	Causal intervention analysis
Acceptance baseline performance	Clinical appropriate evaluation
Basic feasibility check	Expert reasoning alignment scoring
Necessary but insufficient	Required for deployment

Table 5.1: The proposed two-tier validation framework separates traditional performance metrics (Tier 1) from reasoning-based evaluation (Tier 2), highlighting that accuracy alone is not enough for clinical deployment, and that reasoning appropriateness must be validated to ensure patient safety and trust.

5.2.2 Regulatory Alignment

The FDA and international regulatory bodies increasingly require transparency regarding how AI models reason, not only what they predict. Our framework supports regulatory requirements by:

- Producing full audit trails linking evidence to deployment decisions,
- Allowing expert judgment and appropriateness scoring,
- Documenting failure modes and model limitations,
- Supporting traceability under 21 CFR Part 11 compliance through timestamped event logging and version control.

5.2.3 Clinical Workflow Integration

By grounding explanations in anatomical concepts rather than technical artifacts (e.g., gradients, attention weights), the framework aligns explanations with clinicians' normative reasoning. Expert feedback preferred causal intervention analysis (100% usability approval) over correlation-based methods, indicating that clinicians prioritize assessing whether reasoning is correct over which pixels were activated.

5.3 Broader Impact on Medical AI

5.3.1 Generalization Beyond Stroke

While demonstrated on stroke detection, the causal intervention methodology generalizes broadly to medical imaging tasks with established diagnostic criteria. The core principle, assessing feature dependencies against clinical guidelines through systematic perturbation, applies directly to:

- Lung nodule detection: Evaluate dependence on nodule characteristics (spiculation, size, density) versus spurious anatomical position patterns
- Diabetic retinopathy screening: Assess reliance on lesion features (microaneurysms, exudates) versus image quality artifacts

- Breast cancer detection: Verify dependence on lesion morphology and surrounding tissue context versus scanner-specific patterns

5.3.2 Advancing Trustworthy AI

The results demonstrate that accuracy alone is insufficient for safety, that causal reasoning is essential for detecting flawed decision pathways, and that validation must be grounded in domain knowledge and supported by irreplaceable expert evaluation to guide responsible deployment decisions; importantly, because spurious reasoning often reflects algorithmic bias embedded in training data, causal intervention provides a mechanism for early detection and correction of biased dependencies before deployment.

5.4 Contributions to Medical AI Safety

This research advances medical AI safety through three contributions. First, demonstrating the necessity of validating reasoning beyond accuracy. A 92% Dice model rejected by 100% of experts proves that performance-first validation is inadequate, challenging practices that prioritize statistical metrics. Second, establishing the effectiveness of causal interventions to uncover hidden risks. The framework identified 56.2% spurious dependencies that five XAI methods missed, indicating that correlation-based explanations are insufficient. Third, providing a scalable framework with audit processes. Documentation from causal analysis through regulatory reports enables validation, protecting patient safety while preventing dangerous shortcuts.

5.5 Limitations and Future Work

Limitations:

Several limitations warrant acknowledgment. First, vascular territory definitions used simplified geometric heuristics rather than patient-specific anatomy. Second, the simulated clinician panel, while based on documented expertise profiles, requires validation with real clinical experts across multiple institutions. Third, the analysis focused exclusively on DWI sequences, missing multimodal integration. Fourth, the perturbation used binary masking; alternative strategies, such as synthetic inpainting or counterfactual generation, may reveal additional insights. Finally, causal intervention requires multiple forward passes, increasing computational cost compared to single-pass correlation methods.

Future Work:

Research directions include: multi-center clinical validation studies with real expert panels, extension to multimodal imaging (DWI + ADC + FLAIR + perfusion), development of reasoning-aware training objectives that penalize spurious dependencies, integration of uncertainty quantification, automated appropriateness prediction to scale evaluation, and continuous post-deployment monitoring to detect distributional shifts. These extensions would strengthen clinical utility and facilitate broader adoption across medical imaging applications.

6: CONCLUSION

6.1 Summary of Contributions

This research addressed a critical gap in medical AI validation: traditional methods cannot distinguish between models that achieve high accuracy through clinically appropriate reasoning and those that rely on spurious correlations. We developed a causal intervention framework that systematically evaluates whether stroke detection models rely on diagnostically valid features by perturbing anatomically defined vascular territories and quantifying the resulting changes in prediction. Our results demonstrate that accuracy-based validation is insufficient. The analyzed case achieved a 92% Dice coefficient yet demonstrated fundamentally unsafe reasoning with 56.2% dependence on contralateral hemisphere features, a spurious correlation unanimously rejected by all five clinical experts (100% detection rate, 1.0/5 agreement). Critically, five widely used XAI methods (GradCAM, Integrated Gradients, Occlusion, LIME, Attention Rollout) failed to detect this dangerous dependency. The framework successfully decoupled accuracy from reasoning quality, enabling rejection of high-performing but unsafe models. Three contributions emerge: demonstrating that reasoning validation is essential for clinical safety, establishing that causal intervention detects hidden risks evading correlation-based analysis, and providing a scalable deployment framework that prevents dangerous shortcuts before clinical practice.

6.2 Recommendations

For AI developers, we recommend adopting causal reasoning assessment as a routine validation step, defining domain-specific appropriateness criteria collaboratively with clinicians, and generating complete audit trails documenting reasoning. For clinical institutions, we advise requiring reasoning validation before deployment and implementing continuous monitoring for reasoning drift. For regulators, we recommend mandating evidence that accuracy reflects valid reasoning mechanisms and establishing standards for the appropriateness of reasoning. For researchers, future work should apply causal intervention frameworks to additional imaging tasks, develop reasoning-aware training objectives, automate the detection of spurious correlations, and conduct prospective clinical evaluation studies.

7. REFERENCES

[1] Feigin VL, et al. Global burden of stroke, 1990-2019. Lancet Neurology. 2021;20(10):795-820. https://doi.org/10.1016/S1474-4422(21)00252-0

[2] Virani SS, et al. Heart disease and stroke statistics—2021 update. Circulation. 2021;143(8):e254-e743. https://doi.org/10.1161/CIR.0000000000000950

[3] Saver JL. Time is brain—quantified. Stroke. 2006;37(1):263-266. https://doi.org/10.1161/01.STR.0000196957.55928.ab

[4] Nogueira RG, et al. Thrombectomy 6 to 24 hours after stroke. NEJM. 2018;378(1):11-21. https://doi.org/10.1056/NEJMoa1706442

[5] Albers GW, et al. Thrombectomy for stroke at 6 to 16 hours. NEJM. 2018;378(8):708-718. https://doi.org/10.1056/NEJMoa1713973

[6] Schellinger PD, et al. Role of diffusion MRI for acute ischemic stroke diagnosis. Neurology. 2010;75(2):177-185. https://doi.org/10.1212/WNL.0b013e3181e7c9dd

[7] Fiez JA, et al. Lesion segmentation reliability. Human Brain Mapping. 2000;9(4):192-211. https://doi.org/10.1002/(SICI)1097-0193(2000)9:4

[8] Hernandez Petzsche MR, et al. ISLES 2022 dataset. Scientific Data. 2022;9:762. https://doi.org/10.1038/s41597-022-01875-5

[9] Litjens G, et al. Deep learning in medical image analysis survey. Medical Image Analysis. 2017;42:60-88. https://doi.org/10.1016/j.media.2017.07.005

[10] Oakden-Rayner L, et al. Hidden stratification in medical imaging ML. ACM CHIL. 2020:151-159. https://doi.org/10.1145/3368555.3384468

[11] Zech JR, et al. Variable generalization in pneumonia detection. PLoS Medicine. 2018;15(11):e1002683. https://doi.org/10.1371/journal.pmed.1002683

[12] Selvaraju RR, et al. Grad-CAM: Visual explanations from deep networks. ICCV. 2017:618-626. https://openaccess.thecvf.com/content_iccv_2017/html/Selvaraju_Grad-CAM_Visual_Explanations_ICCV_2017_paper.html

[13] Zeiler MD, Fergus R. Visualizing CNNs. ECCV. 2014:818-833. https://doi.org/10.1007/978-3-319-10590-1_53

[14] FDA. Clinical Decision Support Software Guidance. 2022. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-decision-support-software

[15] Ronneberger O, et al. U-Net for biomedical segmentation. MICCAI. 2015:234-241. https://doi.org/10.1007/978-3-319-24574-4_28

[16] Cai CJ, et al. Onboarding needs for human-AI collaboration. ACM CSCW. 2019;3:1-24. https://doi.org/10.1145/3359206

[17] Lipton ZC. The mythos of model interpretability. Queue. 2018;16(3):31-57. https://doi.org/10.1145/3236386.3241340

[18] Pearl J, Mackenzie D. The Book of Why. Basic Books. 2018. https://www.basicbooks.com/titles/judea-pearl/the-book-of-why/9780465097609/

[19] Schölkopf B, et al. Toward causal representation learning. IEEE Proceedings. 2021;109(5):612-634. https://doi.org/10.1109/JPROC.2021.3058954

[20] Yang Q, et al. Unbox the black-box for medical XAI. Information Fusion. 2022;77:29-52. https://doi.org/10.1016/j.inffus.2021.07.016

[21] Sendak MP, et al. Human-centered clinical decision support. JAMIA. 2020. https://doi.org/10.1093/jamia/ocaa120

[22] Tonekaboni S, et al. What clinicians want from XAI. MLHC. 2019. https://proceedings.mlr.press/v106/tonekaboni19a.html

[23] Geirhos R, et al. Shortcut learning in deep neural networks. Nature Machine Intelligence. 2020;2(11):665-673. https://doi.org/10.1038/s42256-020-00257-z

[24] https://www.kaggle.com/datasets/orvile/isles-2022-brain-stoke-dataset/data