Explainable AI for Early Stroke Detection

Causal and Clinician-Centered Interpretation of Brain Imaging

PRESENTED BY

DAVID T.A BLEMANO
DENNIS OWUSU

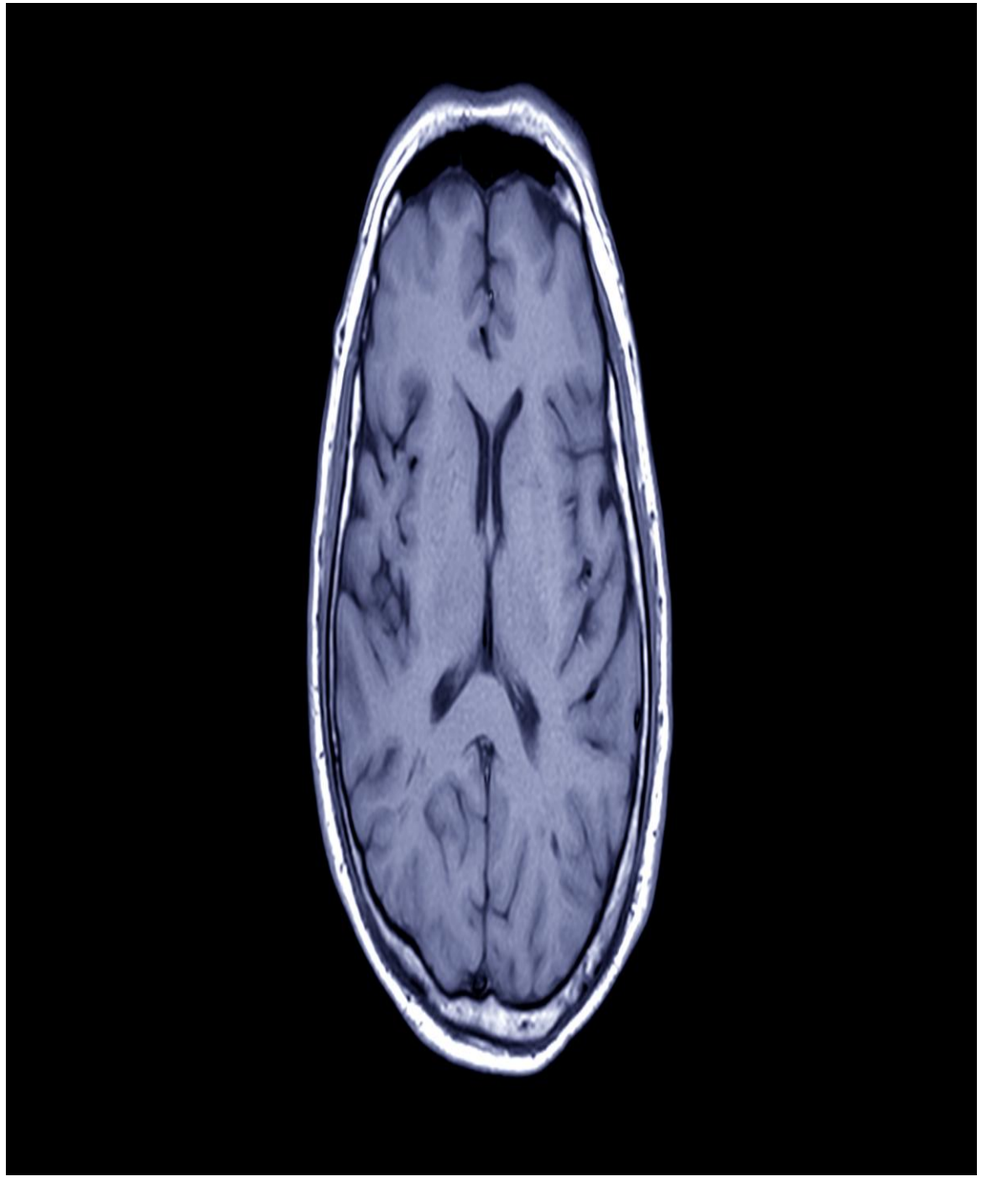
SUPERVISOR

DR. GUY HEMBROFF

The Urgency of Stroke Detection

Acute ischemic stroke is a leading cause of death and disability, affecting 13.7 million people annually. In the US, 795,000 strokes occur yearly, costing over $53 billion.

Prompt treatment is crucial: "time is brain", with nearly two million brain cells dying each minute. Mechanical thrombectomy offers hope, but quick, accurate identification is essential.



Challenges in Diagnosis & AI

1

DWI Limitations

Diffusion-weighted imaging (DWI) is key, but interpreting scans requires expertise. Clinicians must differentiate true infarcts from other bright spots, and lesion outlining varies among experts (Dice coefficient 0.55-0.75).

2

AI "Black Box" Problem

Deep learning models achieve high accuracy but can rely on irrelevant features (e.g., portable machine labels, contralateral symmetry) instead of true pathology, leading to unreliable diagnoses.

3

XAI's Shortcomings

Traditional Explainable AI (XAI) methods show statistical links, not causal relationships. They answer "where" the model looks, not "why" it reasons like a clinician. This gap poses risks for patient safety and clinical trust.

Research Problem & Objectives

Evaluation of current stroke detection models emphasizes accuracy metrics such as Dice, sensitivity, and specificity, but overlooks whether the AI model applies correct clinical reasoning in stroke detection. This gap risks deploying unreliable models that fail on atypical cases or new data. This study aims to bridge this gap by:

Causal Testing Framework

Develop a framework to verify if models use clinically meaningful features, not just correlations, using brain blood-supply regions.

Clinician Evaluation

Create an evaluation plan for medical specialists to assess model reasoning against accepted clinical practice.

Regulatory Compliance

Build a pipeline with an audit trail for transparent, accountable clinical deployment.

Causal Inference Concepts in Medical AI

1 **What is Causal Inference?**

Causal inference distinguishes between mere associations and true cause-effect relationships.

2 **Pearl's Framework Contribution**

Pearl's framework formalizes interventions (do-operator) that fix variables to assess outcome changes, breaking ordinary correlations

3 **Structural Causal Models in Use**

Structural causal models represent variables and dependencies as directed graphs and equations.

4 **Application in Medical Imaging**

Applying causal interventions to medical imaging can isolate the real impact of specific anatomical features on model decisions.

5 **Addressing Bias in Algorithms**

This approach helps detect shortcut learning or unintended biases, like reliance on scanner artifacts or contralateral symmetry.

6 **Advantages of Causally Grounded Models**

Causally grounded models tend to generalize better, are more interpretable, and are more trustworthy.

Related Works

Deep Learning for Stroke Detection

- Early methods used hand-crafted features; performance limited due to lesion variability.
- CNNs improved results; U-Net became standard for segmentation using skip-connections.
- Shift from 2D slices to 3D CNNs captured anatomical context; Dice scores ~0.59–0.70.
- Top ISLES-2022 models reach ~0.75 Dice, but results vary due to protocol differences and annotation uncertainty.
- Clinical adoption remains low: requires interpretability, stability, regulatory validation, and integration into workflows.

Explainable AI in Medical Imaging

- Gradient-based methods (e.g., Grad-CAM, Integrated Gradients) highlight influential regions but mainly show associations, not causation.
- Perturbation/occlusion methods test importance more directly.
- LIME/SHAP and attention maps have stability and scalability limits for 3D imaging.
- Major limitation: explanations rarely prove causal dependence.

Causal Inference for Reliable Systems

- Causal frameworks (Pearl) evaluate effects of interventions (do(X)) and reduce confounding.
- Causal methods improve robustness and generalize across environments.
- Few imaging studies analyze case-level causal evidence behind predictions.

Clinician-Centered Design

- Explanations must align with **clinical reasoning** and workflow needs.
- Trust requires calibrated reliability and transparent decision traces.
- Real deployment demands clinician feedback, usability studies, and audit trails.

Research Gaps

- Accuracy ≠ correct reasoning.
- XAI methods show correlations, not causal justification.
- Minimal work links causal evidence to diagnostic criteria and clinician review.

This work introduces a causal intervention framework that tests whether model decisions depend on truly diagnostic regions, grounded in vascular territories with structured clinician evaluation.

Explainability in Medical Imaging: Existing Techniques

GRADIENT-BASED METHODS

GradCAM and Integrated Gradients highlight pixels influencing predictions.
These methods offer quick yet correlation-based insights.

PERTURBATION TECHNIQUES

Techniques like Occlusion mask regions to determine their impact on model output.
They provide more direct importance measures for explainability.

SURROGATE MODELS

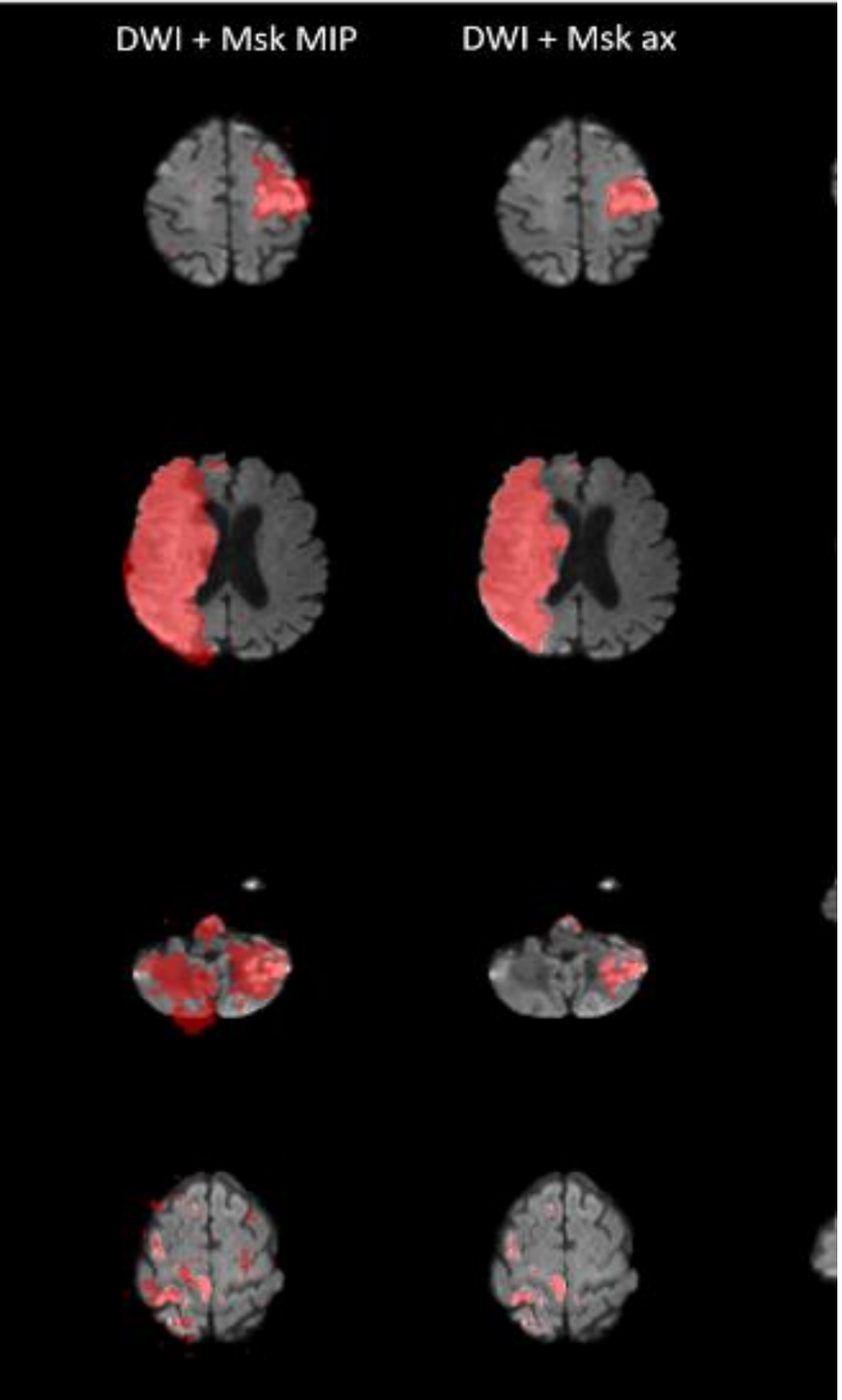
Methods such as LIME and SHAP estimate feature contributions locally.
These methods might be unstable and less practical for 3D scans.

ATTENTION MAPS

Attention maps reveal areas where the model focuses.
They may indicate statistical associations rather than true causal drivers.

LIMITATIONS OF CURRENT TECHNIQUES

Existing techniques reveal where models focus during decision-making, but they cannot confirm whether this attention aligns with clinical reasoning or rests on valid diagnostic evidence—highlighting a critical gap in producing clinically meaningful explanations.



Dataset Overview and Population Justification

Dataset Utilized

ISLES 2022 dataset: 250 multi-center diffusion-weighted MRI scans with manual infarct masks.

Coverage of Stroke Subtypes

Broad spectrum covering diverse stroke subtypes, lesion sizes ranging from 0.1 mL to over 100 mL.

Data Diversity

Collected across multiple countries and hospitals, reflecting varied imaging protocols and quality levels.

Balanced Dataset Splits

Stratified splits ensure balanced representation across lesion sizes and vascular territories, enabling robust model application.

Training Set: 175 (70%)
Validation Set: 37 (14.2%)
Test Set: 38 (15.2%)

Generalization and Evaluation

Facilitates exploration of model generalization and mitigation of potential demographic or scanner bias.

Methodology Overview

MRI Preprocessing

MRI scans undergo standardized preprocessing, including resizing to 80×80×80, skull stripping, and normalization, ensuring consistent input quality for effective model training and evaluation.

Stratified Split

A stratified split is employed, dividing the dataset into 70% training, 14.8% validation, and 15.2% test sets, promoting fair evaluation and reducing bias in model performance assessment.

3D U-Net Model

The architecture of the 3D U-Net employs an encoder-decoder structure with skip connections, batch normalization, ReLU activation, and max pooling to reduce resolution, and is trained using Dice loss and the Adam optimizer for efficient voxel-level segmentation of stroke lesions.

Causal Interventions

Causal interventions are conducted across vascular territories (Middle (MCA), Anterior (ACA), Posterior (PCA) Cerebral Arteries, suspected lesion, and contralateral hemisphere) to assess whether model predictions rely on accurate anatomical evidence, thereby enhancing the validity of the reasoning behind the segmentation results.

Baseline XAI Methods

Baseline explainable AI (XAI) methods, including Grad-CAM, IG, Occlusion, LIME, and Attention Rollout, are utilized for comparative analysis, offering insights into model interpretability and assessment reliability.

Training

Training was performed using Dice loss optimized via the Adam optimizer, configured with a learning rate of 0.001.

Early stopping was implemented to prevent overfitting during training.

Data augmentation (including flips, rotations, and elastic deformations) simulated in vivo variations to enrich the training dataset.

Framework Design and Simulated Clinicians

CAUSAL INTERVENTION FRAMEWORK DESIGN

To assess clinically appropriate reasoning, brain volumes were divided into five vascular territories: Middle (MCA), Anterior (ACA), Posterior (PCA) Cerebral Arteries, suspected lesion, and contralateral hemisphere.

Counterfactual Test

Counterfactual intervention replaces voxel intensities in each territory separately with noise reflecting brain-tissue statistics.

Causal Contribution

The difference in Dice score measured the causal effect of each region.

Clinical Assessment

Performance reduction >60% in lesion area indicated relevant features. Moderate effects (20-80%) in affected territories were appropriate in context. Minimal change (<20%) in the opposite hemisphere confirmed correct localization. Strong drop (>40%) in contralateral suggested spurious reliance.

SIMULATED CLINICIAN EVALUATION

Composition

- A simulated panel was created using established clinical scoring guidelines.
- Simulated panel modeled on five experts: three neuroradiologists and two stroke neurologists with varied experience levels.
- Profiles reflect different interpretation styles, from conservative to aggressive.

Focus Areas Evaluated

- The structured questionnaire included questions assessing the anatomical correctness of segmentations and the clinical relevance by vascular territory.
- The study evaluated the diagnostic importance of features such as DWI hyperintensity, lesion shape, and clinical context.

Deployment Recommendations Given

- The panelists provided feedback on the system's readiness for deployment, offering recommendations: Approve, Conditional, or Reject.
- These evaluations are crucial for refining deployment strategies.

Experimental Results

MODEL PERFORMANCE

0.6780	0.6108	38	0.9232
Best Val Dice	Test Mean Dice	Test Cases	Analyzed Case

Causal Intervention Analysis

⚠ Spurious Reasoning Detected

56.2% dependence on **contralateral hemisphere** — clinically inappropriate spatial heuristic

✓ Appropriate Features

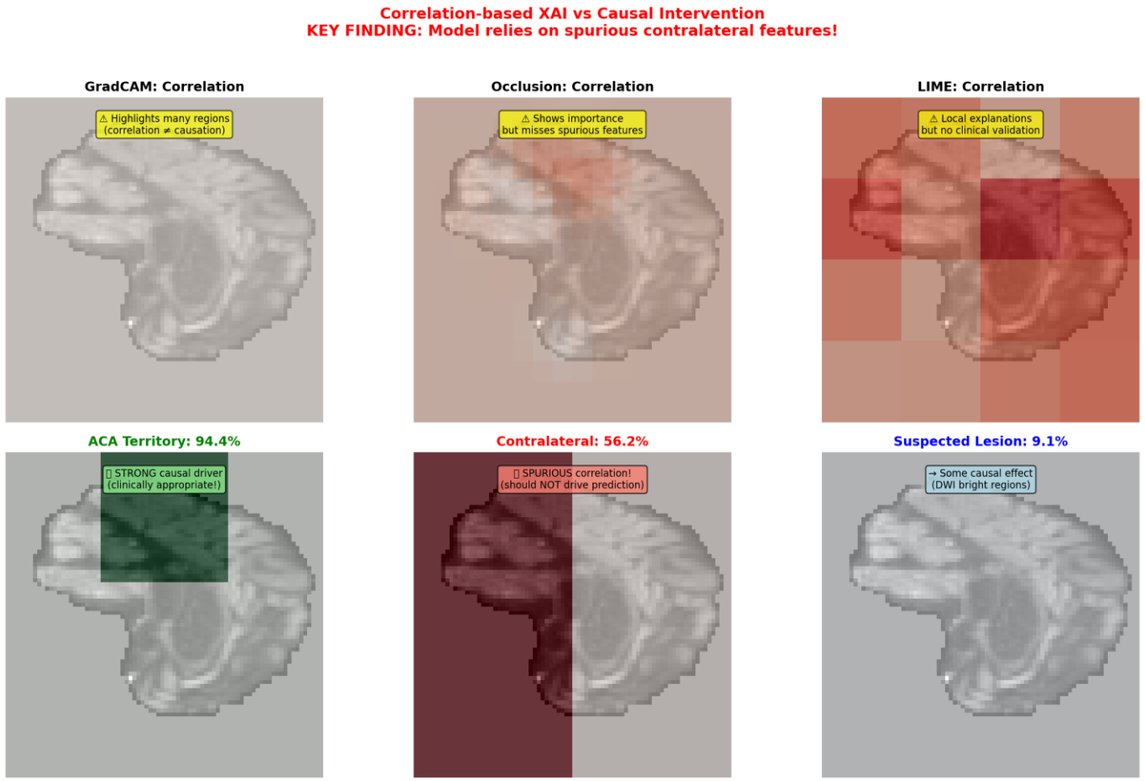
94.4% from **ACA territory** — valid vascular recognition

Minimal Lesion Reliance

Only **9.1%** from **DWI lesion** (expected >60% for appropriate reasoning)

Clinical Evaluation

5/5 experts rejected (100% spurious detection) • Agreement: 1.0/5 • Trust: 1.16/5



Key Finding: All correlation-based XAI methods failed to detect spurious dependency

Discussions

Critical Discovery

1. Reasoning Validation Necessity

92% Dice model rejected by 100% of experts proves performance-first validation inadequate

2. Causal Intervention Success

Identified 56.2% spurious dependency that five XAI methods completely missed

3. Scalable Framework

Complete audit system enables validation protecting patient safety

The Accuracy – Trustworthiness Gap

Models achieve high metrics while relying on inappropriate reasoning

- **Evidence:** 92% Dice yet 100% clinical rejection
- **Cause:** 56.2% spurious reliance, only 9.1% on lesion
- **Risk:** False confidence; would fail on bilateral strokes

Bridging Statistical Performance and Clinical Trustworthiness

Paradigm Shift: From performance-first to reasoning-centered validation, prioritizing patient safety.

Tier-Two Validation Required

- Tier 1: Performance thresholds
- Tier 2: Reasoning appropriateness

Conclusion and Recommendations

KEY ACHIEVEMENT

This study shows that reasoning validation is not optional and must complement performance metrics in clinical AI. Unanimous expert rejection validates domain-knowledge-grounded assessment, bridging the accuracy-trustworthiness gap.

RECOMMENDATIONS

AI RESEARCHERS

- Two-tier validation
- Causal frameworks

CLINICAL INSTITUTIONS

- Review protocols
- Continuous monitoring

REGULATORY BODIES

- Updated guidance
- Audit requirements

STUDY LIMITATIONS

- Simplified vascular territories (binary masks)
- Single imaging modality (DWI only)
- Limited clinical evaluation (5 experts, simulated)
- Single dataset (ISLES 2022, n=250)
- Computational overhead of causal interventions

FUTURE RESEARCH DIRECTIONS

- Multi-center clinical validation trials
- Multimodal imaging integration (DWI, FLAIR, PWI)
- Reasoning-aware training objectives
- Automated appropriateness prediction
- Cross-domain application (lung, cardiac imaging)

FINAL REMARK: ONLY BY VALIDATING HOW AND WHY MODELS PREDICT, THEN CAN WE BUILD AI SYSTEMS WORTHY OF CLINICAL TRUST AND ENSURE PATIENT SAFETY

Cited References

Feigin VL, et al. Lancet Neurology. 2021;20(10):795-820. Link

Virani SS, et al. Circulation. 2021;143(8):e254-e743. Link

Saver JL. Stroke. 2006;37(1):263-266. Link

Nogueira RG, et al. NEJM. 2018;378(1):11-21. Link

Albers GW, et al. NEJM. 2018;378(8):708-718. Link

Schellinger PD, et al. Neurology. 2010;75(2):177-185. Link

Fiez JA, et al. Human Brain Mapping. 2000;9(4):192-211. Link

Hernandez Petzsche MR, et al. Scientific Data. 2022;9:762. Link

Selvaraju RR, et al. ICCV. 2017:618-626. Link

Zeiler MD, Fergus R. ECCV. 2014:818-833. Link

FDA. Clinical Decision Support Software Guidance. 2022. Link

Pearl J, Mackenzie D. The Book of Why. Basic Books. 2018. Link

Schölkopf B, et al. IEEE Proceedings. 2021;109(5):612-634. Link

Yang Q, et al. Information Fusion. 2022;77:29-52. Link

Sendak MP, et al. JAMIA. 2020. Link

Tonekaboni S, et al. MLHC. 2019. Link

Geirhos R, et al. Nature Machine Intelligence. 2020;2(11):665-673. Link

Kaggle. ISLES 2022 Brain Stroke Dataset. Link

Complete bibliography available in project documentation

THANK YOU

For Questions/Comments/Suggestions/Collaboration:

Dennis Owusu
denniso@mtu.edu
www.github.com/RoyalDennis
www.linkedin.com/in/dennis-owusu

David T.A Blemano
dtbleman@mtu.edu
www.github.com/David5-cyber
www.linkedin.com/in/david-blemano