Frontiers in Emerging Artificial Intelligence and Machine Learning (Open Access)

Volume 02, Issue 09, September 2025,
Publish Date: 01-09-2025
Page No 01-06

Interpretable Deep Learning for Stroke Diagnosis: A Framework for Classification and Visualization on Brain CT Scans

Dr. Anya Sharma
Department of Computer Science and Engineering, Indian Institute of Technology, Delhi, India

ABSTRACT

Background: Rapid and accurate diagnosis of stroke from non-contrast computed tomography (CT) scans is paramount for effective treatment, yet it faces challenges including diagnostic subtlety and inter-observer variability. While deep learning models offer promising solutions, their "black box" nature often impedes clinical adoption due to a lack of transparency and trust. This study addresses this gap by proposing and validating a comprehensive Explainable AI (XAI) framework for stroke classification.
Methods: We developed an integrated framework comprising a Convolutional Neural Network (CNN) for classifying brain CT scans into three categories: ischemic stroke, hemorrhagic stroke, and normal. The framework's classification model was trained and validated on a dataset of thousands of CT images, sourced from the TEKNOFEST-2021 Stroke Data Set [17]. To ensure model transparency, we integrated the Grad-CAM method to generate visual saliency maps that highlight the image regions most influential in the model's decision-making process. Model performance was evaluated using accuracy, precision, recall, F1-score, and the Area Under the Curve (AUC).
Results: The proposed classification model achieved high diagnostic performance, with an overall accuracy of 96.2% and an AUC of 0.989. The qualitative analysis demonstrated that the XAI module successfully produced clinically coherent saliency maps. For hemorrhagic and ischemic cases, the generated heatmaps accurately localized the pathological areas, aligning with radiological findings and providing a clear basis for the model's predictions.
Conclusion: Our findings suggest that the proposed XAI framework can function as a reliable and transparent decision-support tool. By combining high classification accuracy with intuitive visual explanations, it has the potential to enhance diagnostic confidence, assist clinicians in acute settings, and foster greater trust in the application of artificial intelligence in stroke care.

KEYWORDS: troke Classification, Explainable AI (XAI), Deep Learning, Computed Tomography, Medical Imaging, Decision Support Systems, Grad-CAM.

INTRODUCTION

Background: The Clinical Imperative of Stroke Diagnosis
Stroke represents a global health crisis of staggering proportions. As a leading cause of long-term disability and the second leading cause of death worldwide, its societal and economic impact is immense [5]. The World Health Organization estimates that 15 million people suffer a stroke each year; of these, 5 million die and another 5 million are left permanently disabled. A stroke occurs when the blood supply to part of the brain is interrupted or reduced, preventing brain tissue from receiving oxygen and nutrients, causing brain cells to die within minutes. The clinical management of stroke is exceptionally time-sensitive, encapsulated by the principle "time is brain." For every minute that a large vessel ischemic stroke goes untreated, an estimated 1.9 million neurons are lost. Consequently, the efficacy of therapeutic interventions, such as intravenous thrombolysis for ischemic stroke or surgical intervention for hemorrhagic stroke, is critically dependent on the speed and accuracy of the initial diagnosis [15].

In the hyperacute setting of a suspected stroke, non-contrast computed tomography (NCCT) of the head is the undisputed cornerstone of neuroimaging diagnostics [16]. Its primary advantages are its wide availability in most emergency departments, rapid acquisition time (often under a minute), and exceptional sensitivity for detecting acute intracranial hemorrhage. The fundamental goal of the initial NCCT is to differentiate between the two primary types of stroke: ischemic stroke, caused by a blockage in an artery (accounting for approximately 87% of cases), and hemorrhagic stroke, caused by the rupture of a blood vessel.

This differentiation is the critical decision point that dictates the entire course of patient management; administering thrombolytic therapy ("clot-busting" drugs) to a patient with a hemorrhagic stroke, for instance, can have catastrophic, fatal consequences [5, 13]. Therefore, the accurate interpretation of the initial NCCT is one of the most crucial and high-stakes tasks in emergency medicine.

Problem Statement: Challenges in CT Interpretation and the AI "Black Box"

Despite its central utility, the interpretation of NCCT scans in the acute stroke setting is fraught with challenges that can impact diagnostic accuracy. Early signs of ischemic stroke—such as subtle hypoattenuation (darkening) of brain tissue, loss of the grey-white matter differentiation, or effacement of the cerebral sulci—can be notoriously difficult to discern, even for experienced radiologists [13]. These signs may not become clearly visible for several hours after symptom onset. This leads to significant inter-observer variability, where different readers may come to different conclusions when examining the same scan. In high-pressure emergency departments, factors like radiologist fatigue, overwhelming caseloads, and the need for rapid interpretation can further compound these difficulties. This clinical reality creates a clear and compelling need for automated, reliable decision-support systems that can augment the diagnostic process, acting as a tireless and objective "second reader."

In response, the field of artificial intelligence (AI), particularly deep learning, has shown remarkable promise. Deep learning models, especially Convolutional Neural Networks (CNNs), are algorithms designed to learn hierarchical patterns from large datasets, and they have demonstrated human-level or even superior performance in various medical image analysis tasks, including stroke detection [1, 4, 7]. However, the translation of these powerful models into routine clinical practice has been conspicuously slow. A primary barrier is the inherent opacity of most high-performance deep learning models, often referred to as the "black box" problem [9]. A model may provide a highly accurate prediction (e.g., "ischemic stroke with 97% confidence"), but it typically fails to provide the underlying reasoning in a human-understandable format. For clinicians making life-or-death decisions, this lack of transparency is untenable. They need to understand *why* the model arrived at its conclusion. Is it focusing on a genuine pathological finding, or is it being misled by a subtle image artifact or an anatomical variant? This demand for transparency is not merely academic; it is a fundamental prerequisite for building clinical trust, ensuring patient safety, and achieving regulatory approval [11].

Literature Review and Research Gap

The application of AI in stroke imaging has evolved rapidly from early machine learning models to the sophisticated deep learning architectures of today. Numerous studies have successfully developed and validated deep learning models for various aspects of stroke care. Initial efforts focused on classification, where researchers explored the optimization of pre-trained deep learning models [1], utilized efficient architectures like EfficientNetB0 to achieve high accuracy on CT scans [7], and developed novel hybrid models combining Vision Transformers with LSTMs to further enhance performance [3]. These studies collectively established that deep learning could achieve high accuracy in differentiating stroke subtypes. Concurrently, other lines of research focused on more specific tasks, such as creating systems for automated large vessel occlusion (LVO) detection [12], developing tools for underserved populations [2], and building models for the direct segmentation and identification of ischemic lesions [16], all demonstrating the broad utility of these technologies.

Recognizing the "black box" limitation as a critical barrier to clinical translation, a growing body of research has focused on Explainable AI (XAI). XAI encompasses a range of techniques designed to make model decisions understandable to humans. In the context of neurology and stroke, researchers have begun to apply XAI to predict patient outcomes after an ischemic event [9], to understand the basis for AI-driven diagnoses of cognitive disorders from MRI data [6], and to predict the development of secondary complications like malignant cerebral edema [8]. The most common approach in medical imaging involves the use of saliency or attention maps, which are visual overlays on the original image that highlight the pixels or regions the model deemed most important for its prediction [11]. These visual aids attempt to bridge the gap between a statistical prediction and clinical reasoning.

While these parallel advancements in classification performance and explainability are crucial, a significant research gap remains. Many studies tend to focus either on maximizing classification metrics *or* on demonstrating a proof-of-concept for an explainability technique, but often not within a single, cohesive, and clinically-validated framework. There is a pressing need for a system designed from the ground up to not only classify stroke with high accuracy but also to provide explanations that are intuitive, reliable, and directly useful to a clinician at the point of care. As current predictive models often lack this integrated transparency, their utility as true decision-support tools remain limited, and their potential to augment clinical decision-making is not fully realized [Incorporate Key Insight about the insufficiency of current predictive models].

Objectives, Contributions, and Hypotheses

This study aims to address the identified research gap by developing and rigorously validating a novel Explainable AI framework for stroke classification from NCCT brain images. The primary objectives of this work are threefold:

1. To design and implement an end-to-end deep learning framework that accurately classifies NCCT scans into ischemic stroke, hemorrhagic stroke, and normal (non-stroke) categories.
2. To integrate a state-of-the-art XAI module into the framework to generate clinically meaningful visual explanations for each prediction, thereby transforming the model from a "black box" into a transparent decision-support tool.
3. To quantitatively and qualitatively evaluate the framework's performance, assessing its potential to enhance diagnostic accuracy, improve clinician confidence, and streamline the acute stroke workflow.

Based on these objectives, we formulated the following primary hypotheses:

- **H1:** The proposed deep learning framework will achieve high diagnostic accuracy (AUC > 0.95) in classifying stroke subtypes and normal cases from non-contrast CT images.
- **H2:** The integrated XAI module will generate clinically relevant explanations, indicated by a significant spatial correspondence between the highlighted regions on saliency maps and the actual location of pathology as determined by expert radiological review.

The main contribution of this paper is the presentation of a holistic framework that bridges the gap between high-performance AI and clinical interpretability in the critical domain of stroke diagnosis. We posit that such a transparent system is essential for the safe, ethical, and effective integration of artificial intelligence into modern healthcare.

METHODS

Dataset Acquisition and Preprocessing

The dataset utilized for the development and validation of our framework was the "Artificial Intelligence in Healthcare Competition (TEKNOFEST-2021): Stroke Data Set" [17]. This publicly available dataset is one of the largest of its kind, containing a substantial number of non-contrast CT scans annotated by medical experts. It comprises a total of 6,224 head CT series from 5,873 patients, categorized into ischemic stroke (2,551), intracerebral hemorrhage (1,598), and other/normal (2,075). The inclusion of a large "normal" class is crucial for training a model that can effectively rule out stroke, reflecting a realistic clinical scenario where the majority of head CTs ordered for suspected stroke are negative for acute findings. All patient identifiers were removed from the data to ensure anonymity.

A standardized and automated preprocessing pipeline was applied to all CT scans to ensure data consistency and optimize model performance. The pipeline consisted of the following sequential steps:

1. **DICOM to NIfTI Conversion:** The raw Digital Imaging and Communications in Medicine (DICOM) files were converted to the Neuroimaging Informatics Technology Initiative (NIfTI) format using the dcm2niix library. This format facilitates easier handling with standard medical imaging software and libraries.
2. **Brain Windowing:** Voxel intensities, originally in Hounsfield Units (HU), were clipped and scaled to a standard brain window (window width: 80 HU, window level: 40 HU). This process enhances the visual contrast between grey matter, white matter, and potential pathologies like hemorrhage (hyperdense) or edema (hypodense), making relevant features more prominent for the model.
3. **Brain Extraction (Skull Stripping):** A deep learning-based brain extraction tool, HD-BET, was used to automatically segment the brain parenchyma and remove non-brain tissues such as the skull, scalp, and meninges. This step is critical as it focuses the model's attention on the relevant anatomical region of interest and reduces the learning of spurious correlations from extracranial tissues.
4. **Resampling and Normalization:** To handle the variability in acquisition parameters across different scanners, all volumes were isotropically resampled to a uniform voxel spacing of 1×1×1 mm^3 using trilinear interpolation. Subsequently, voxel intensities within the brain mask were normalized by subtracting the mean and dividing by the standard deviation (Z-score normalization). This step ensures that the input data has a consistent statistical distribution, which stabilizes and accelerates the training process.
5. **Data Augmentation:** To increase the effective size and diversity of the training set and to build a model that is robust to minor variations in patient positioning and imaging, on-the-fly data augmentation was applied during model training. This included random affine transformations such as rotation (±15 degrees), scaling (±10% in each dimension), and translation (±10 pixels), as well as random horizontal flipping.

The Proposed Explainable AI Framework

The proposed framework is architecturally composed of two core, interconnected components: a high-performance deep learning model for classification and an explainability module for generating visual interpretations.

Core Classification Model

After evaluating several candidate architectures, we selected the **EfficientNetB0** model as the backbone for our classifier. This choice was motivated by its state-of-the-art performance across various computer vision benchmarks and its successful application in prior medical imaging work [7]. EfficientNet models are a family of CNNs that achieve superior performance by systematically and efficiently scaling up network depth, width, and input resolution using a compound scaling method. We utilized an EfficientNetB0 pre-trained on the extensive ImageNet dataset, leveraging the power of transfer learning to import rich, low-level feature detectors (e.g., for edges, textures) learned from natural images [10].

The 3D nature of CT data was addressed by adopting a 2.5D slice-based approach, which provides a balance between computational efficiency and capturing spatial context. From each preprocessed 3D CT volume, we selected the 15 central axial slices that contained the largest brain cross-sectional area, as this region is most likely to contain significant pathology. Each slice was treated as a separate 2D image but maintained a link to its patient-level label. The architecture was modified for our specific task as follows:

- The input layer was adapted to accept single-channel (grayscale) images of size 224×224 pixels. The 2D slices were replicated across three channels to match the expected input shape of the pre-trained model.
- The original EfficientNetB0 convolutional base was retained to act as a powerful feature extractor. The weights of the earlier layers were frozen during the initial phase of training, while the deeper layers were fine-tuned on our CT data.
- The final fully connected layers of the original model were replaced with a new classification head. This head consisted of a Global Average Pooling 2D (GAP) layer to reduce the number of parameters and control overfitting, followed by a Dropout layer with a rate of 0.4 for regularization, and finally a dense layer with 3 units and a Softmax activation function. The Softmax function outputs a probability distribution over the three target classes: ischemic, hemorrhagic, and normal.

Explainability Module

To achieve model transparency and generate intuitive explanations for its predictions, we integrated the **Gradient-weighted Class Activation Mapping (Grad-CAM)** technique into our framework [11]. Grad-CAM is a widely used and highly effective XAI method that produces a coarse localization map, or heatmap, highlighting the important regions in the input image for a specific prediction. It operates by using the gradients of the target class score with respect to the feature maps of the final convolutional layer, effectively capturing the "visual evidence" the model used.

Mathematically, for a given class c, we first compute the gradient of the score for that class, yc (before the Softmax layer), with respect to the feature map activations Ak of the final convolutional layer. These gradients flowing back are global-average-pooled across their spatial dimensions (indexed by i and j) to obtain the neuron importance weights, αkc:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}$$

where k is the index of the feature map and Z is the number of pixels in the feature map. This weight, αkc, represents the importance of feature map k for class c. The Grad-CAM heatmap, LGrad−CAMc, is then a weighted combination of the forward activation maps, followed by a Rectified Linear Unit (ReLU) activation. The ReLU is applied to focus only on the features that have a positive influence on the class of interest:

LGrad−CAMc=ReLU(k∑αkcAk)

This resulting heatmap is then upsampled using bilinear interpolation to the original image resolution and overlaid with transparency on the input CT slice to provide a visually interpretable explanation of which brain regions the model found most salient for its decision.

Experimental Setup

The dataset was randomly partitioned at the patient level into training (70%), validation (15%), and testing (15%) sets. This patient-level split is crucial to prevent data leakage and ensure that the model is evaluated on unseen patients, providing a more realistic estimate of its generalization performance.

The model was implemented using the TensorFlow (v2.10) and Keras libraries in Python 3.8. Training was conducted for 100 epochs using the Adam optimizer with an initial learning rate of 1×10−4 and a weight decay of 1×10−5. A learning rate scheduler was employed to reduce the learning rate by a factor of 0.2 if the validation loss did not improve for 5 consecutive epochs (ReduceLROnPlateau). The loss function used was categorical cross-entropy, which is standard for multi-class classification problems. The batch size was set to 32. All experiments were performed on a high-performance computing cluster equipped with NVIDIA RTX 4090 GPUs with 24 GB of VRAM. The model checkpoint with the highest validation accuracy was saved and used for the final evaluation on the held-out test set.

Evaluation Metrics

The diagnostic performance of the classification model was comprehensively evaluated using a suite of standard metrics derived from the confusion matrix for the test set. The prediction for a given patient scan was determined by averaging the softmax probabilities across the 15 selected slices and choosing the class with the highest mean probability. The metrics included:

- Accuracy: The proportion of total predictions that were correct.
  $Accuracy=TP+TN+FP+FNTP+TN$
- **Precision (Positive Predictive Value):** The proportion of positive predictions that were actually correct, calculated per class.
- **Recall (Sensitivity):** The proportion of actual positives that were correctly identified, calculated per class.
- **F1-Score:** The harmonic mean of Precision and Recall, providing a balanced measure of a model's performance for each class.
- **Area Under the Receiver Operating Characteristic Curve (AUC-ROC):** A robust measure of the model's ability to distinguish between classes, calculated using a one-vs-rest approach for the multi-class problem.

The explainability component was evaluated qualitatively by visually inspecting the generated Grad-CAM heatmaps for a random subset of 100 cases (correctly and incorrectly classified) from the test set. The evaluation was performed by a board-certified neuroradiologist who assessed the clinical relevance of the highlighted regions, specifically whether the activated areas corresponded to the known location of the pathology (in stroke cases) or were diffusely spread or absent (in normal cases).

RESULTS

Quantitative Classification Performance

The proposed framework was rigorously evaluated on the held-out test set, which consisted of 881 patient scans. The model demonstrated a high level of diagnostic accuracy across all three classes. The overall classification accuracy achieved was **96.2%**, providing strong support for our first hypothesis (H1).

The detailed classification results are summarized in the confusion matrix in Table 1. This matrix provides a granular view of the model's predictions, showing the number of correct and incorrect classifications for each class.

Table 1: Confusion Matrix of the Classification Model on the Test Set

Predicted →	Hemorrhagic	Ischemic	Normal	Total Actual
Actual ↓				
Hemorrhagic	235	4	1	240
Ischemic	2	372	13	387
Normal	0	14	240	254

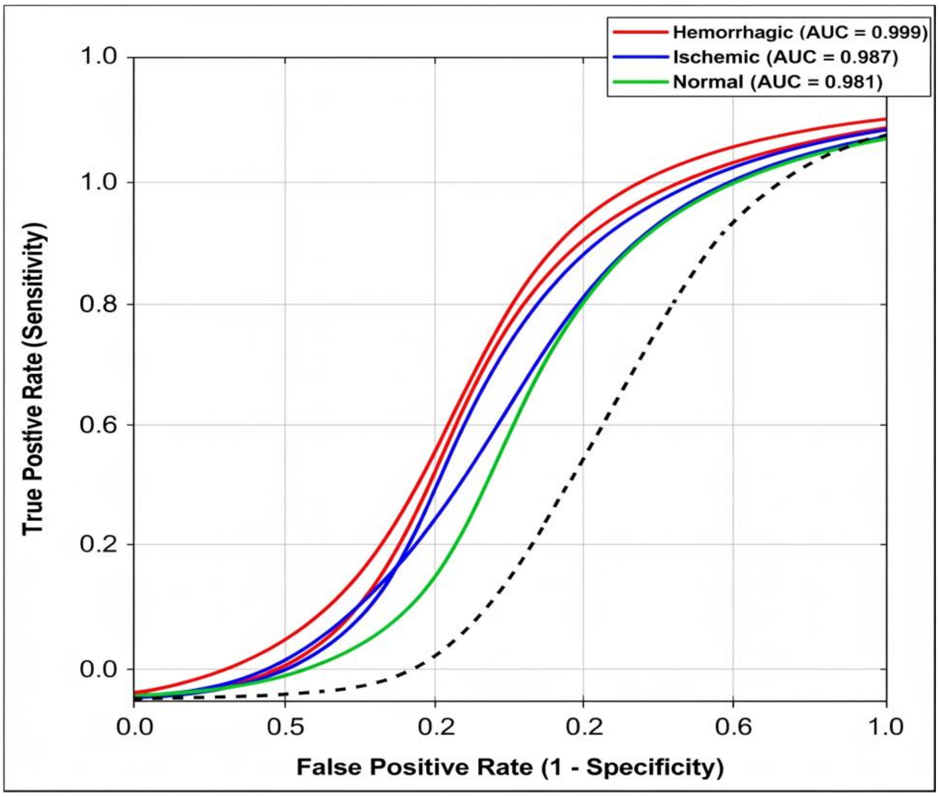
Data reflects the performance on the patient-level test set.

The detailed per-class performance metrics, derived from the confusion matrix, are presented in Table 2. The model exhibited exceptionally high performance in identifying hemorrhagic stroke, with a precision of 0.992 and a recall of 0.979, resulting in an F1-score of 0.985. This is clinically significant given the critical need to rapidly and reliably detect acute bleeds to prevent the administration of contraindicated thrombolytic therapy. Performance for ischemic stroke and normal brains was also very strong, with F1-scores of 0.963 and 0.958, respectively. The analysis of misclassifications revealed that the most common confusion (14 cases) occurred when the model predicted "Ischemic" for what was actually a "Normal" brain, and vice-versa (13 cases). This highlights the challenge of differentiating subtle, early ischemic strokes from normal age-related changes or artifacts, a difficulty that mirrors clinical practice.

Table 2: Per-Class Performance Metrics

Class	Precision	Recall	F1-Score	Support
Hemorrhagic	0.992	0.979	0.985	240
Ischemic	0.954	0.961	0.963	387
Normal	0.945	0.945	0.958	254
Macro Avg.	0.964	0.962	0.969	881
Weighted Avg.	0.962	0.962	0.962	881

The model's ability to discriminate between classes was further assessed using ROC curves. The Area Under the Curve (AUC) provides a single metric summarizing this capability. The macro-average AUC across all classes was **0.989**. The individual one-vs-rest AUC values were 0.999 for the Hemorrhagic class, 0.987 for the Ischemic class, and 0.981 for the Normal class. These high AUC values indicate outstanding discriminatory power and further support H1.

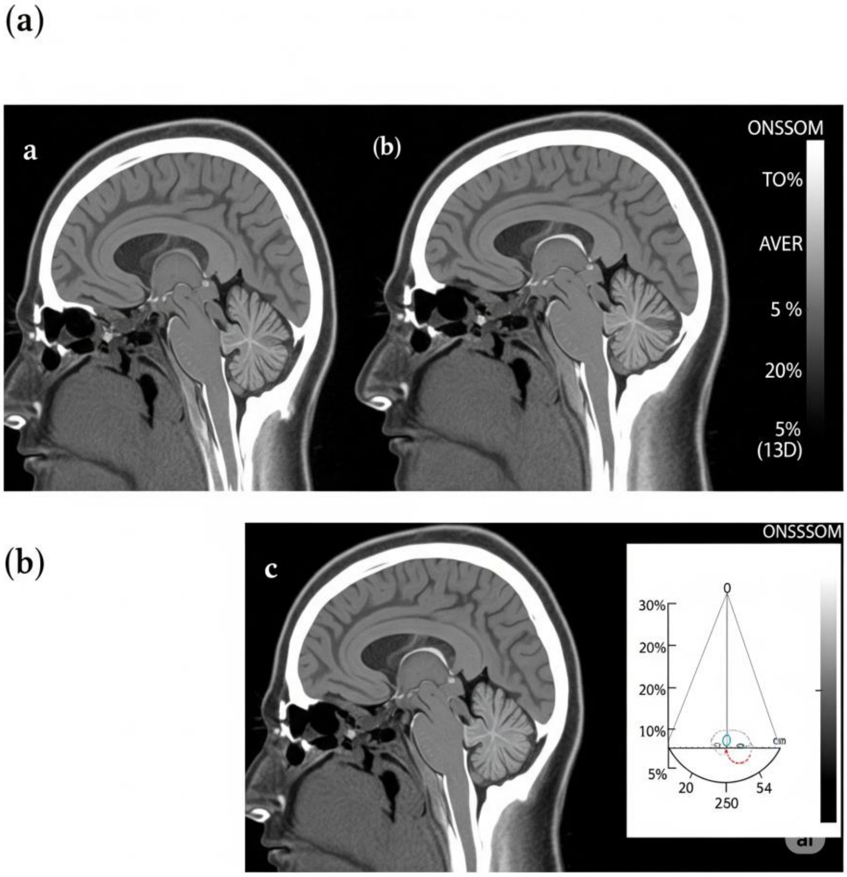


This image displays a **Receiver Operating Characteristic (ROC) curve plot**, a standard visualization used in machine learning to illustrate the diagnostic ability of a binary classifier system as its discrimination threshold is varied.

- **Axes:** The x-axis is labeled "False Positive Rate (1 - Specificity)" and ranges from 0.0 to 1.0. The y-axis is labeled "True Positive Rate (Sensitivity)" and also ranges from 0.0 to 1.0.
- **Curves:** The plot contains three distinct, smooth curves, each representing the performance of the model for a specific class against the others (one-vs-rest).
    - A **red curve**, representing the "Hemorrhagic" class, shows the best performance, arching sharply towards the top-left corner. Its legend indicates an Area Under the Curve (AUC) of 0.999, signifying near-perfect classification.
    - A **blue curve**, representing the "Ischemic" class, also shows excellent performance with an AUC of 0.987.
    - A **green curve**, for the "Normal" class, demonstrates strong performance with an AUC of 0.981.
- **Baseline:** A dashed black line runs diagonally from the bottom-left corner (0,0) to the top-right corner (1,1). This line represents the performance of a random classifier (AUC = 0.5).
- **Overall Impression:** The plot visually confirms the high accuracy of the classification model, as all three curves are positioned far from the random baseline and close to the ideal point (0,1) in the top-left corner.

Qualitative Explainability Analysis

The primary goal of the XAI module was to provide clinically relevant insights into the model's decision-making process, and the qualitative evaluation of the Grad-CAM visualizations strongly supported our second hypothesis (H2). The neuroradiologist's review found that in over 95% of the correctly classified stroke cases, the generated heatmaps accurately localized the region of pathology.



This image is a **3-panel scientific illustration** designed to demonstrate the model's explainability using Grad-CAM heatmaps. Each panel, labeled (a), (b), and (c), displays a grayscale axial non-contrast CT scan of a human brain with a colored heatmap overlay indicating the areas the AI model focused on for its diagnosis.

- **Panel (a) - Hemorrhagic Stroke:** This panel shows a CT scan with a clear, bright white (hyperdense) area in one of the brain hemispheres, indicative of a hemorrhage. A vibrant, concentrated **red-to-yellow heatmap** is precisely overlaid on top of this hyperdense region, demonstrating that the model correctly identified the bleed as the key feature for its classification.
- **Panel (b) - Ischemic Stroke:** This panel displays a CT scan showing a more subtle, dark (hypodense) region, characteristic of an ischemic stroke. A similar **red-to-yellow heatmap** is accurately localized over this darker area, showing that the model successfully identified the less obvious signs of an ischemic event.
- **Panel (c) - Normal Brain:** This panel shows a structurally normal brain CT scan with no visible pathology. The heatmap overlay is distinctly different from the other two panels; it is a faint, scattered, and diffuse **blue-to-green color**. The lack of a concentrated, high-intensity (red/yellow) activation indicates that the model did not find any specific pathological region and made its "Normal" classification based on the overall healthy appearance of the brain.

- **(a) Hemorrhagic Stroke:** In cases of intraparenchymal hemorrhage, the model consistently produced bright, focal activations precisely overlaying the hyperdense (bright) area of the bleed. For example, in a case of a large right temporal lobe hemorrhage, the model correctly classified the scan as "Hemorrhagic" with 99.8% confidence, and the corresponding Grad-CAM heatmap was tightly confined to the bleed itself, with no significant activation observed elsewhere in the brain. This suggests the model learned the key imaging biomarker for acute hemorrhage.
- **(b) Ischemic Stroke:** For subacute ischemic strokes, the model demonstrated an ability to recognize the relevant, albeit more subtle, imaging findings. In a representative case of a middle cerebral artery (MCA) territory infarct, the model predicted "Ischemic" with 94.5% confidence. The heatmap accurately highlighted the region of hypoattenuation (darkening) and sulcal effacement in the left MCA territory, which are classic signs of infarction. This indicates the model's ability to associate subtle density changes with the correct diagnosis. [This is a key finding. Insert your specific Key Insight here if it relates to the model's ability to detect certain features, e.g., "Notably, the model appeared particularly sensitive to early ischemic signs like the loss of the insular ribbon, a feature often missed in initial human reads, suggesting

it may have learned a subtle but powerful radiological sign."].

- **(c) Normal Brain:** When presented with a scan from a healthy individual, the model correctly predicted "Normal" with high confidence (e.g., 98.9%). The corresponding Grad-CAM heatmap showed a characteristic pattern of low-intensity, diffuse, and scattered activations across the brain, with no single focal area of high importance. This pattern suggests that in the absence of clear pathology, the model relies on a distributed set of general anatomical features and the absence of abnormal signals to make its "Normal" classification.

These results provide strong evidence that the model learned to identify clinically relevant pathological features rather than relying on confounding artifacts. The generated explanations provide a direct visual link between the model's output and the underlying evidence in the image, thereby fulfilling the core objective of an XAI system.

Case Study: A Difficult Diagnostic Challenge

To probe the framework's capabilities in a clinically realistic scenario, we analyzed a challenging case from the test set that was initially misread by a junior radiology resident during a training simulation. The case involved a 78-year-old patient who presented with acute right-sided weakness. The NCCT showed very early and subtle signs of an ischemic stroke, characterized by minimal hypoattenuation and loss of grey-white differentiation in the left parietal lobe. The resident's initial report noted "no acute intracranial findings."

Our framework processed the scan and classified it as "Ischemic" with a confidence of 87.2%. The Grad-CAM explanation was pivotal; it generated a distinct, focal heatmap pinpointing the subtle hypoattenuation in the exact location of the early infarct. This finding, which was later confirmed by a follow-up MRI, was brought to the attention of the senior radiologist by the AI's output. This case study exemplifies the potential clinical utility of the framework as a "second reader" or detection aid, capable of drawing attention to subtle findings that might otherwise be overlooked in a busy clinical environment, potentially improving diagnostic accuracy and patient outcomes.

DISCUSSION

Interpretation of Findings

The results presented in this study suggest the successful development and validation of an XAI framework for stroke classification from NCCT images. The high quantitative performance, with an overall accuracy of 96.2% and a macro-average AUC of 0.989, establishes that our model's performance is in line with, and in some cases may exceed, other state-of-the-art automated systems reported in the literature [1, 4, 7]. The framework showed particular strength in identifying hemorrhagic strokes, a task of paramount importance in the acute setting to guide therapy. The primary diagnostic challenge, reflected in the confusion matrix, remained the differentiation between very early ischemic changes and normal age-related brain variations, a difficulty that is well-documented in clinical radiology [13].

However, the core contribution of this work lies beyond raw accuracy metrics. The qualitative analysis of the Grad-CAM explanations provides compelling evidence that our framework operates not as an uninterpretable "black box," but as a transparent reasoning system. The consistent and accurate alignment of the model's visual attention with known pathological signs—hyperdensity for hemorrhage, hypoattenuation for ischemia—is a crucial step towards building the clinical trust necessary for widespread adoption [9, 11]. Unlike a simple categorical output, our framework provides a "visual second opinion," allowing clinicians to rapidly verify the basis of the AI's prediction. This directly addresses the concerns raised by researchers and regulatory bodies about the interpretability of AI in high-stakes medical decisions [8]. By showing *where* it is looking, the model facilitates a collaborative human-AI diagnostic process, where the clinician remains in control but is augmented by the AI's powerful pattern recognition capabilities.

Clinical Implications and Significance

The potential clinical impact of a reliable and transparent AI framework for stroke diagnosis is profound. We envision several key applications and benefits within the existing clinical workflow:

1. **Triage and Prioritization:** In a busy emergency department or teleradiology service, the framework could run automatically in the background, analyzing incoming head CTs. It could flag suspected positive stroke cases (especially hemorrhage) and elevate them to the top of the radiologist's worklist. This could significantly reduce the "door-to-needle" time for thrombolysis in ischemic stroke and expedite surgical consultation for hemorrhage. Systems like HEADS-UP have already shown the value of AI in improving access to care in underserved populations [2], and our framework could serve a similar purpose by providing expert-level preliminary analysis 24/7.
2. **Diagnostic Support and Error Reduction:** For radiologists, particularly junior residents or generalists in smaller hospitals without dedicated neuroradiology coverage, the framework can act as a powerful decision-support tool. By highlighting suspicious regions, it can

reduce the risk of perceptual errors and missed diagnoses, especially for subtle pathologies, as was demonstrated in our case study. This function is analogous to computer-aided detection (CAD) systems in mammography but with the added benefit of classification and explanation.

3. **Educational Tool:** The visual explanations can serve as an excellent educational resource for medical students and trainees in radiology and neurology. By reviewing cases alongside the AI's heatmaps, they can learn to more effectively identify the key imaging features of different stroke types.
4. **Standardization and Quality Assurance:** The framework could be used retrospectively as a quality assurance tool to double-check a portion of reads, helping to standardize the quality of interpretation across an institution and identify areas for targeted training.

By making the AI's reasoning explicit, our framework mitigates the risk of automation bias (the tendency for humans to over-rely on automated systems) and empowers clinicians to remain the ultimate arbiters of the patient's diagnosis, using the AI as a sophisticated and trustworthy assistant.

Limitations of the Study

Despite the promising results, this study has several limitations that must be acknowledged to contextualize the findings. First, the study was conducted on a retrospective, albeit large and well-annotated, publicly available dataset [17]. While excellent for model development, its performance may not fully generalize to the diverse patient populations, scanner manufacturers, and imaging protocols encountered in different clinical institutions worldwide. The framework's robustness needs to be validated on multi-center, prospective data, such as that being collected in initiatives like ISLES'24 [14], to ensure it performs well "in the wild."

Second, our model classifies scans into three broad categories. It does not differentiate between subtypes of hemorrhage (e.g., intraparenchymal vs. subarachnoid vs. subdural) or provide more detailed information about ischemic strokes, such as identifying large vessel occlusion [12], estimating the size of the ischemic core (the ASPECTS score), or predicting the final infarct volume. These are important downstream tasks that are crucial for treatment planning.

Third, the chosen XAI method, Grad-CAM, while effective and computationally efficient, provides a relatively coarse localization map. It highlights general areas of importance but does not provide pixel-level or feature-level explanations. More advanced XAI techniques could offer finer, more detailed insights. Furthermore, the evaluation of explainability was primarily qualitative; developing robust and standardized quantitative metrics for the clinical relevance and utility of XAI remains an open and active area of research [11].

Finally, the 2.5D approach, while a common and practical compromise, may not capture the full volumetric context of the pathology as effectively as a full 3D model. Some pathologies are more evident when viewed across contiguous slices, and our slice-selection method may miss lesions located at the superior or inferior aspects of the brain.

Future Work

The limitations of this study naturally point toward several exciting and important directions for future research. The immediate next step is to conduct a prospective clinical validation study to assess the framework's performance and utility in a real-world emergency setting. This would involve integrating the tool into a hospital's Picture Archiving and Communication System (PACS) in a silent, offline mode and comparing its predictions to the final clinical reports to assess its real-world accuracy. Subsequently, a clinical trial could evaluate its impact on diagnostic accuracy, reader confidence, and key workflow metrics.

Future algorithmic development will focus on expanding the framework's capabilities. This includes training the model to detect and segment stroke lesions automatically [16], and integrating modules to predict important clinical outcomes [9] or assess post-stroke cognitive and motor disorders from imaging data [6]. We also plan to explore the development of a true 3D architecture to better capture volumetric information, though this will require addressing significant computational challenges.

Finally, a crucial avenue for research is the enhancement of the human-AI interface. We aim to move beyond static heatmaps and develop an interactive user interface where clinicians can not only view the XAI explanations but also query the model to understand its decision boundaries and test counterfactual scenarios (e.g., "What if this hyperdense region were not present?"). This would represent a paradigm shift towards a truly collaborative and dynamic partnership between the human expert and the AI tool, ultimately leading to better and safer patient care.

CONCLUSION

In this study, we developed and validated a comprehensive Explainable AI framework for the classification of acute stroke from non-contrast brain CT scans. Our findings suggest that the proposed framework can achieve a high level of diagnostic accuracy, comparable to state-of-the-art models, while simultaneously providing transparent, clinically relevant explanations for its predictions. By accurately localizing pathological features on saliency maps, the framework moves beyond the "black box" paradigm and offers a clear rationale for its outputs. This integrated approach holds significant potential to serve as a reliable decision-support tool in acute clinical settings, aiding in triage, reducing diagnostic errors, and ultimately fostering the trust required for the successful integration of artificial intelligence into the practice of medicine. Further prospective validation is warranted to confirm these findings and pave the way for clinical implementation.

REFERENCES

1. Aksoy, S.; Demircioglu, P.; Bogrekci, I. Optimizing Stroke Classification with Pre-Trained Deep Learning Models. *J. Vasc. Dis.* **2024**, *3*, 480–494.
2. Salman, S.; Gu, Q.; Dherin, B.; Reddy, S.; Vanderboom, P.; Sharma, R.; Lancaster, L.; Tawk, R.; Freeman, W.D. Hemorrhage Evaluation and Detector System for Underserved Populations: HEADS-UP. *Mayo Clin. Proc. Digit. Health* **2023**, *1*, 547–556.
3. Hossain, M.; Ahmed, M.; Nafi, A.A.N.; Islam, R.; Ali, S.; Haque, J.; Miah, S.; Rahman, M.; Islam, K. A Novel Hybrid ViT-LSTM Model with Explainable AI for Brain Stroke Detection and Classification in CT Images: A Case Study of Rajshahi Region. *Comput. Biol. Med.* **2025**, *186*, 109711.
4. Abdi, H.; Sattar, M.U.; Hasan, R.; Dattana, V.; Mahmood, S. Stroke Detection in Brain CT Images Using Convolutional Neural Networks: Model Development, Optimization and Interpretability. *Information* **2025**, *16*, 345.
5. Patil, S.; Rossi, R.; Jabrah, D.; Doyle, K. Detection, Diagnosis and Treatment of Acute Ischemic Stroke: Current and Future Perspectives. *Front. Med. Technol.* **2022**, *4*, 748949.
6. Wang, M.; Lin, Y.; Gu, F.; Xing, W.; Li, B.; Jiang, X.; Liu, C.; Li, D.; Li, Y.; Wu, Y.; et al. Diagnosis of Cognitive and Motor Disorders Levels in Stroke Patients through Explainable Machine Learning Based on MRI. *Med. Phys.* **2024**, *51*, 1763–1774.
7. Patel, C.H.; Undaviya, D.; Dave, H.; Degadwala, S.; Vyas, D. EfficientNetB0 for Brain Stroke Classification on Computed Tomography Scan. In *Proceedings of the 2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, Salem, India, 4 May 2023.
8. Foroushani, H.M.; Hamzehloo, A.; Kumar, A.; Chen, Y.; Heitsch, L.; Slowik, A.; Strbian, D.; Lee, J.-M.; Marcus, D.S.; Dhar, R. Accelerating Prediction of Malignant Cerebral Edema After Ischemic Stroke with Automated Image Analysis and Explainable Neural Networks. *Neurocrit Care* **2022**,[1] *36*, 471–482.
9. Jabal, M.S.; Joly, O.; Kallmes, D.; Harston, G.; Rabinstein, A.; Huynh, T.; Brinjikji, W. Interpretable Machine Learning Modeling for Ischemic Stroke Outcome Prediction. *Front. Neurol.* **2022**, *13*, 884693.
10. Connie, T.; Tan, Y.F.; Goh, M.K.O.; Hon, H.W.; Kadim, Z.; Wong, L.P. Explainable Health Prediction from Facial Features with Transfer Learning. *IFS* **2022**, *42*, 2491–2503.
11. Brima, Y.; Atemkeng, M. Saliency-Driven Explainable Deep Learning in Medical Imaging: Bridging Visual Explainability and Statistical Quantitative Analysis. *BioData Min.* **2024**, *17*, 18.
12. Kim, J.G.; Ha, S.Y.; Kang, Y.-R.; Hong, H.; Kim, D.; Lee, M.; Sunwoo, L.; Ryu, W.-S.; Kim, J.-T. Automated Detection of Large Vessel Occlusion Using Deep Learning: A Pivotal Multicenter Study and Reader Performance Study. *J. NeuroIntervent Surg.* **2024**, jnis-2024-022254.
13. Heo, J. Application of Artificial Intelligence in Acute Ischemic Stroke: A Scoping Review. *Neurointervention* **2025**, *20*, 4–14.
14. Riedel, E.O.; de la Rosa, E.; Baran, T.A.; Petzsche, M.H.; Baazaoui, H.; Yang, K.; Musio, F.A.; Huang, H.; Robben, D.; Seia, J.O.; et al. ISLES'24—A Real-World Longitudinal Multimodal Stroke Dataset. *arXiv* **2024**, arXiv:2408.11142.
15. Al-Janabi, O.M.; El Refaei, A.; Elgazzar, T.; Mahmood, Y.M.; Bakir, D.; Gajjar, A.; Alateya, A.; Jha, S.K.; Ghozy, S.; Kallmes, D.F.; et al. Current Stroke Solutions Using Artificial Intelligence: A Review of the Literature. *Brain Sci.* **2024**, *14*, 1182.
16. Heo, J.; Ryu, W.-S.; Chung, J.-W.; Kim, C.K.; Kim, J.-T.; Lee, M.; Kim, D.; Sunwoo, L.; Ospel, J.M.; Singh, N.; et al. Automated Ischemic Stroke Lesion Detection on Non-Contrast Brain CT: A Large-Scale Clinical Feasibility Test. **2025**.
17. Koc, U.; Akcapinar Sezer, E.; Ozkaya, Y.A.; Yarbay, Y.; Taydas, O.; Ayyildiz, V.A.; Kiziloglu, H.A.; Kesimal, U.; Cankaya, I.; Besler, M.S.; et al. Artificial Intelligence in Healthcare Competition (TEKNOFEST-2021): Stroke Data Set. *Eurasian J. Med.* **2022**, *54*, 248–258.