GMR
The Original

AI DRIVEN NEUROVISION-FM: A SELF-SUPERVISED MULTIMODAL FOUNDATION MODEL FOR GENERALIZABLE BRAIN MRI ANALYSIS

Dr. Kommu Naveen[1*], Dr. S China Venkateswarlu[2], B. Jyothi[3], Dr. Ashok Nayak[4], Divya Manasa Malisetty[5], Dr. Battula Balnarsaiah[6], Dr. Bhaludra R Nadh Singh[7]

[1*]Professor, Department of Computer Science Engineering- AIML, Kasireddy Narayanareddy College of Engineering and Research, Hyderabad-501513, Telangana State, India. naveenkarunya@gmail.com https://orcid.org/0000-0001-6459-1549
[2]Professor of Computer Science Engineering, Faculty of Information Technology & Engineering, Gopal Narayan Singh University, Sasaram, Bihar, schina.venkateswarlu@gnsu.ac.in, cvenkateswarlus@gmail.com
[3]Associate Professor, Computer Science and Machine Learning (CSM), St Peter's Engineering College Hyderabad, India - 500100, bjyothi815@gmail.com, Orcid.org: 0000-0002-9342-6603
[4]Associate Professor, Department of Electronics and Communication Engineering, Marri Laxman Reddy Institute of Technology and Management (UGC Autonomous), Dundigal, Hyderabad-500043, Telangana, India. ashoka405@gmail.com
[5]Assistant Professor, Department of Electronics and Communications Engineering, TKR College of Engineering and Technology, TKRCET, Medbowli, Meerpet, Saroornagar, Hyderabad -500097, Telangana, India. divya.java3213@gmail.com
[6]Associate Professor, Department of AIML, Nalla Malla Reddy Engineering College, Divyanagar, Narapally, Hyderabad, Telangana State, India – 500088, battulabalu@gmail.com
[7]Professor, Computer Science Engineering, Neil Gogte Institute of Technology, Hyderabad, Telangana, India-500 088. brn.singh@gmail.com

ABSTRACT

Brain magnetic resonance imaging (MRI) analysis is increasingly supported by artificial intelligence, yet many existing models remain dependent on task-specific supervised learning, large expert-labeled datasets, single-modality inputs, and narrowly sampled institutional cohorts. These limitations can reduce generalization across scanners, acquisition protocols, hospitals, and patient populations. This paper proposes NeuroVision-FM, a self-supervised multimodal foundation-model framework for generalizable brain MRI analysis. The framework combines self-supervised representation learning, CNN-based local feature extraction, Transformer-based global contextual modeling, multimodal clinical feature encoding, cross-modal attention, adaptive feature fusion, uncertainty estimation, and explainable artificial intelligence. The design supports downstream classification, lesion localization/segmentation, and risk-oriented assessment while maintaining patient-level separation during evaluation. A rigorous evaluation protocol is specified using public brain-MRI datasets, patient-level partitioning, external validation, class-balanced metrics, calibration analysis, robustness testing, ablation studies, and statistical confidence intervals. Explainability is assessed through image-region attribution and structured-feature importance, with clinical interpretability reviewed against expert-annotated or clinically meaningful regions where feasible. The manuscript deliberately distinguishes the proposed architecture from a fully established clinical foundation model: numerical performance values and figures must be generated from real experiments and must not be represented as measured evidence until validated. NeuroVision-FM is positioned as a reproducible research framework for studying whether self-supervised multimodal representations can improve cross-domain robustness, label efficiency, interpretability, and transferability in brain MRI analysis. Brain magnetic resonance imaging (MRI) analysis has increasingly benefited from artificial intelligence (AI) and deep learning (DL); however, many existing approaches remain constrained by task-specific supervised learning, limited expert annotations, single-modality inputs, institutional domain bias, and insufficient interpretability. These limitations can substantially affect the reliability and generalizability of AI systems when they are transferred across scanners, acquisition protocols, healthcare institutions, and patient populations. To address these challenges, this study proposes NeuroVision-FM, a self-supervised multimodal foundation-model framework for generalizable brain MRI analysis. The proposed architecture integrates self-supervised representation learning with complementary convolutional neural network (CNN) and Vision Transformer (ViT) encoders to capture local anatomical patterns and long-range contextual dependencies. A multimodal representation layer incorporates structured clinical information and, where available, electronic health records and medical text through cross-modal attention and adaptive feature fusion. The framework further incorporates uncertainty estimation and explainable artificial intelligence (XAI) to improve transparency, reliability, and clinical interpretability.

The proposed learning pipeline consists of patient-level data partitioning, MRI quality control, modality harmonization, preprocessing and augmentation, self-supervised pretraining, multimodal representation alignment, adaptive fusion, task-specific fine-tuning, uncertainty estimation, and explanation generation. The framework is designed to support multiple downstream applications, including brain-tumor classification, lesion localization and segmentation, disease characterization, and risk-oriented prediction. Evaluation is designed around rigorous patient-level validation using accuracy, precision, sensitivity, specificity, F1-score, ROC-AUC, PR-AUC, calibration error, computational efficiency, and robustness under distribution shift. For segmentation tasks, Dice similarity coefficient, Jaccard index/IoU, Hausdorff distance, and sensitivity are considered. Statistical confidence intervals, ablation experiments, external validation,

missing-modality analysis, and subgroup evaluation are incorporated to assess the reliability and transferability of the learned representations.

A central objective of NeuroVision-FM is to investigate whether self-supervised multimodal learning can reduce dependence on expensive annotations while improving cross-domain generalization and clinical interpretability. The proposed framework is therefore positioned not merely as a task-specific diagnostic classifier but as a transferable representation-learning architecture capable of adaptation to diverse brain-MRI analysis tasks. Importantly, all quantitative performance values and clinical claims must be established through reproducible experiments on independent datasets and should not be inferred from the proposed architecture alone. The study provides a methodological foundation for developing privacy-conscious, generalizable, interpretable, and clinically responsible medical foundation models for next-generation neuroimaging applications.

KEYWORDS: Brain MRI; Foundation Models; Self-Supervised Learning; Multimodal Learning; Vision Transformer; CNN; Explainable AI; Medical Image Analysis; Domain Generalization; Clinical Decision Support; Uncertainty Estimation.

INTRODUCTION

Magnetic resonance imaging provides high-resolution structural and pathological information for neurological assessment, but its interpretation remains challenging because abnormalities can vary in location, size, morphology, intensity, and relationship to surrounding tissue. AI-based image analysis has therefore become an important research direction for automated detection, classification, segmentation, grading, and prognostic assessment. Deep CNNs have demonstrated strong performance in learning hierarchical spatial representations, whereas Transformer architectures provide a complementary mechanism for modeling long-range relationships through attention.

Despite these advances, many published systems remain narrowly optimized for a single task and a single dataset. Medical annotation is costly, and a model trained on one institution may experience substantial performance degradation when applied to another institution because of scanner differences, field strength, acquisition protocols, preprocessing pipelines, demographic variation, and disease prevalence. Furthermore, brain MRI is frequently interpreted together with clinical history, age, symptoms, laboratory findings, radiology reports, or other structured information. An image-only model therefore does not always represent the full clinical decision context.

NeuroVision-FM addresses these challenges by treating self-supervised multimodal representation learning as the central mechanism for building reusable brain-MRI representations. The proposed architecture first learns from unlabeled or weakly labeled MRI data, then integrates image and clinical representations using cross-modal attention and adaptive fusion, and finally adapts the learned representation to downstream tasks. Explainability and uncertainty estimation are included as first-class evaluation components rather than as afterthoughts.

Magnetic resonance imaging (MRI) has become one of the most important non-invasive imaging modalities for the assessment of neurological disorders because of its superior soft-tissue contrast, multiplanar capability, and ability to characterize structural, functional, and pathological alterations within the brain. Brain MRI plays a central role in the detection and assessment of tumors, ischemic and hemorrhagic lesions, neurodegenerative diseases, demyelinating disorders, vascular abnormalities, and other neurological conditions. However, accurate interpretation of brain MRI remains a challenging and highly specialized task. Pathological manifestations may exhibit considerable variability in anatomical location, lesion size, shape, texture, intensity, tissue boundaries, contrast enhancement, and spatial relationship with surrounding structures. Furthermore, MRI examinations frequently contain multiple sequences, such as T1-weighted, contrast-enhanced T1-weighted, T2-weighted, and fluid-attenuated inversion recovery (FLAIR) images, each providing complementary information regarding tissue composition and pathological characteristics. This multimodal nature makes automated brain-MRI analysis both a technically complex and clinically significant research problem.

The rapid advancement of artificial intelligence (AI), machine learning (ML), and deep learning (DL) has significantly transformed biomedical image analysis. Deep neural networks have demonstrated the ability to learn hierarchical representations directly from medical images and have been applied to automated detection, classification, segmentation, grading, prognosis, and treatment-response assessment. In particular, convolutional neural networks (CNNs) have established strong performance in medical image analysis because their convolutional operations efficiently capture local spatial structures, edges, textures, anatomical boundaries, and hierarchical patterns. Architectures based on residual learning and encoder-decoder designs have further improved the ability of deep networks to model complex anatomical structures and pathological regions. However, CNNs primarily rely on local receptive fields and hierarchical convolutional representations, which may limit their ability to efficiently capture long-range relationships between anatomically distant regions of a high-dimensional MRI volume.

Transformer-based architectures have introduced an alternative paradigm for medical image representation learning by employing self-attention mechanisms to model relationships among spatial tokens. Vision Transformers (ViTs) and hybrid CNN-Transformer architectures can capture global contextual dependencies that complement the local inductive biases of CNNs. This capability is particularly relevant to brain MRI, where pathological interpretation may depend not only on the appearance of an individual lesion but also on its relationship to surrounding tissue, anatomical structures, contralateral regions, and other distant regions of the brain. Nevertheless, Transformer-based medical imaging systems may require substantial training data and computational resources, and their effectiveness can be influenced by the quality, diversity, and representativeness of the available training datasets. Consequently, the integration of complementary CNN and Transformer representations represents an important direction for developing robust and transferable brain-MRI analysis systems.

Despite substantial progress in AI-driven medical imaging, a significant proportion of existing brain-MRI models remain designed as task-specific supervised learning systems. Such approaches generally require large quantities of labeled data for a particular diagnostic objective, such as tumor classification or lesion segmentation. The availability of high-quality expert annotations is a major limitation in biomedical research because annotation of MRI examinations requires substantial time, domain expertise, and clinical knowledge. Moreover, annotations can exhibit inter-observer variability, particularly for lesions with ambiguous boundaries or heterogeneous appearance. A model that is optimized exclusively using a relatively small labeled dataset may consequently learn representations that are highly specialized to the characteristics of that dataset rather than generalizable neurological features.

Another important challenge is domain shift and distributional heterogeneity. Brain MRI data collected from different hospitals or research centers can differ substantially because of scanner manufacturer, magnetic-field strength, acquisition sequence, imaging resolution, reconstruction methods, contrast-agent protocols, preprocessing pipelines, patient demographics, disease prevalence, and clinical workflows. A model that achieves excellent performance on an internal test set may therefore experience considerable degradation when evaluated on an independent external cohort. This phenomenon represents one of the principal barriers to translating AI-based neuroimaging research from experimental environments to real-world clinical settings. Consequently, generalization across institutions, scanners, acquisition protocols, and patient populations should be considered a fundamental design objective rather than an optional extension of brain-MRI modeling.

The challenge becomes more pronounced when brain MRI is considered in its actual clinical context. Clinical diagnosis is rarely determined exclusively from image pixels. Radiologists and clinicians may integrate MRI findings with patient age, symptoms, neurological examination, medical history, laboratory measurements, treatment information, radiology reports, and other structured or unstructured clinical evidence. An image-only AI system therefore represents only one component of the information available during clinical decision-making. Multimodal learning provides an opportunity to address this limitation by learning complementary representations from imaging data and associated clinical information. When appropriately designed, multimodal models can potentially identify relationships between imaging biomarkers and clinical characteristics that may not be captured by independent unimodal models.

However, multimodal medical AI introduces additional technical challenges. Different modalities possess different statistical characteristics, dimensionalities, missing-data patterns, semantic structures, and noise distributions. MRI images are high-dimensional spatial signals, whereas clinical variables may consist of categorical, numerical, temporal, or textual information. Consequently, direct concatenation of heterogeneous features may not be sufficient to learn meaningful cross-modal relationships. Advanced representation alignment, cross-modal attention, and adaptive fusion mechanisms are required to determine how information from different modalities should interact and contribute to the final prediction. Furthermore, real-world clinical data are frequently incomplete, meaning that some patients may have MRI data but lack particular clinical variables or textual reports. Therefore, robustness to missing modalities should be explicitly considered during model development and evaluation.

These challenges motivate a transition from conventional task-specific deep-learning architectures toward medical foundation-model paradigms. A medical foundation model can be broadly viewed as a representation-learning system pretrained on large and diverse datasets and subsequently adapted to multiple downstream tasks. In the context of brain MRI, self-supervised learning is particularly attractive because it can exploit large quantities of unlabeled imaging data. Rather than requiring every image to have a manually assigned diagnostic label, self-supervised methods construct learning objectives from the data itself. Depending on the selected strategy, the model can learn representations through masked-image reconstruction, contrastive learning, image transformation prediction, cross-modal alignment, or other pretext objectives. Such pretrained representations may provide a stronger initialization for downstream classification, segmentation, localization, prognosis, and related tasks.

The proposed NeuroVision-FM framework is motivated by this foundation-model paradigm. Rather than learning a narrowly optimized mapping from a single MRI input to a single diagnostic label, NeuroVision-FM is designed to learn a reusable multimodal representation of brain MRI and associated clinical information. The framework combines self-supervised representation learning, CNN-based local feature extraction, Transformer-based global contextual modeling, clinical feature encoding, cross-modal attention, adaptive multimodal feature fusion, uncertainty estimation, and explainable artificial intelligence (XAI). This combination is intended to provide complementary capabilities: CNN components capture local anatomical and pathological structures; Transformer components model long-range contextual relationships; clinical encoders represent non-imaging information; cross-modal attention establishes interactions among heterogeneous representations; and adaptive fusion learns the relative contribution of different information sources.

A central characteristic of NeuroVision-FM is the use of self-supervised pretraining before downstream task adaptation. During the pretraining stage, the model learns generalizable representations from unlabeled or weakly labeled MRI data. The learned representation can subsequently be fine-tuned or adapted for specific tasks using comparatively smaller labeled datasets. This strategy has the potential to improve label efficiency and reduce dependence on extensive manual annotation. Importantly, however, the effectiveness of self-supervised pretraining must be empirically demonstrated through controlled comparisons against supervised baselines, different pretraining objectives, and varying quantities of labeled data.

Another defining component of NeuroVision-FM is multimodal feature fusion. Let (I) represent MRI information and (C) represent structured clinical information. The framework aims to learn a joint representation

[
Z_F = F(Z_I,Z_C),
]

where (Z_I) represents learned imaging features, (Z_C) represents clinical features, and (F(\cdot)) denotes the proposed cross-modal attention and adaptive fusion mechanism. The resulting fused representation can then be transferred to downstream predictive functions corresponding to classification, segmentation, localization, or prognostic analysis. The objective is not merely to concatenate heterogeneous features but to learn clinically meaningful interactions between imaging and contextual information.

Generalizability is another fundamental objective of the proposed framework. Rather than evaluating the model exclusively using randomly selected samples from the same distribution as the training dataset, NeuroVision-FM should be assessed using patient-level partitioning, independent external datasets, cross-domain experiments, missing-modality scenarios, and distribution-shift conditions. Such evaluation can provide stronger evidence regarding whether the learned representations capture clinically meaningful characteristics rather than institution-specific acquisition artifacts. In particular, patient-level partitioning is essential because multiple MRI slices, sequences, examinations, or longitudinal observations from the same patient can otherwise appear simultaneously in training and test sets, producing artificially optimistic performance estimates.

Clinical deployment also requires more than high predictive accuracy. A model may produce highly accurate predictions while relying on spurious correlations, imaging artifacts, acquisition-specific characteristics, or clinically irrelevant regions. Therefore, explainability and uncertainty estimation are incorporated into NeuroVision-FM as first-class components of the proposed research framework. Image-level explanations can be generated using techniques such as Grad-CAM, Grad-CAM++, Integrated Gradients, attention-based attribution, or related methods, while structured clinical features can be interpreted using SHAP or other feature-attribution techniques. These explanations should subsequently be assessed for localization quality, stability, faithfulness, and clinical plausibility rather than being interpreted automatically as causal evidence.

Uncertainty estimation provides an additional layer of reliability assessment. In clinical decision-support applications, an AI system should ideally distinguish between predictions made with high confidence and cases that fall outside its learned distribution or contain ambiguous evidence. Predictive uncertainty can therefore be analyzed together with calibration measures, reliability diagrams, expected calibration error, and selective prediction performance. This enables the research to move beyond the conventional question of whether the model is correct toward the more clinically relevant question of how reliable the model believes its prediction to be.

The proposed research therefore establishes a unified perspective in which self-supervised learning provides scalable representation acquisition, CNNs provide local anatomical modeling, Transformers provide global contextual modeling, multimodal fusion provides clinical context, XAI provides interpretability, and uncertainty estimation provides reliability assessment. These components collectively form the conceptual foundation of NeuroVision-FM.

The principal research hypothesis is that a self-supervised multimodal foundation-model architecture can learn more transferable and clinically meaningful brain-MRI representations than conventional task-specific supervised models, particularly when evaluated under limited-label, cross-domain, external-validation, and missing-modality conditions. The study consequently emphasizes not only predictive performance but also generalization, label efficiency, calibration, robustness, interpretability, computational efficiency, and clinical relevance.

In summary, NeuroVision-FM seeks to bridge the gap between conventional brain-MRI deep-learning systems and more generalizable medical foundation-model architectures. The framework is intended to support adaptation across multiple neurological imaging tasks while incorporating multimodal clinical information and transparent model assessment. Nevertheless, the proposed architecture should be regarded as a research framework until its hypotheses are validated through reproducible experiments. Claims regarding clinical superiority, foundation-model capability, or real-world deployment should be supported by independent datasets, external validation, statistically rigorous comparisons, comprehensive ablation studies, and expert-oriented clinical evaluation. This distinction is essential for ensuring scientific validity and responsible development of AI-based systems for biomedical imaging.

2. MOTIVATION AND RESEARCH HYPOTHESIS

2.1 Motivation

The increasing availability of heterogeneous brain MRI datasets has created an opportunity to develop AI systems that learn generalizable representations rather than isolated task-specific decision functions. Conventional supervised learning approaches typically optimize a model for a particular diagnostic task using a fixed labeled dataset. Although such systems can achieve strong performance under controlled experimental conditions, their effectiveness may deteriorate when exposed to limited labels, unseen scanners, different acquisition protocols, institutional variation, or incomplete clinical information.

The motivation for NeuroVision-FM arises from six fundamental limitations in existing brain-MRI AI systems:

1. Limited availability of expert annotations: Pixel-level and patient-level medical annotations require specialist expertise and considerable time.
2. Task-specific representation learning: Many existing architectures learn features specifically for one classification or segmentation task and may not transfer effectively to other neurological applications.
3. Domain shift: Differences in scanners, field strength, acquisition protocols, reconstruction algorithms, preprocessing, and patient populations can cause substantial distributional changes.
4. Incomplete clinical context: MRI images alone do not necessarily contain all information used by clinicians during diagnosis and prognosis.
5. Limited interpretability: High predictive performance does not guarantee that a model is using clinically meaningful anatomical or pathological evidence.

6. Uncertainty in clinical decision support: A prediction without an estimate of confidence or uncertainty can be problematic when a case is ambiguous or substantially different from the training distribution.
NeuroVision-FM addresses these limitations through a unified architecture that combines self-supervised learning, CNN representation learning, Transformer-based contextual modeling, multimodal clinical fusion, explainable AI, and uncertainty estimation.

2.2 Research Hypothesis

The primary hypothesis is:

H1: A self-supervised multimodal CNN–Transformer foundation-model architecture incorporating clinical information, cross-modal attention, adaptive feature fusion, explainability, and uncertainty estimation can learn more generalizable and clinically meaningful brain-MRI representations than conventional task-specific supervised learning models.

The hypothesis can be decomposed into experimentally testable sub-hypotheses:

- H1a: Self-supervised pretraining improves downstream performance under limited labeled-data conditions.
- H1b: CNN–Transformer hybrid representations outperform either CNN-only or Transformer-only representations.
- H1c: Multimodal image–clinical fusion improves diagnostic performance compared with MRI-only models.
- H1d: Cross-modal attention improves representation quality compared with simple feature concatenation.
- H1e: NeuroVision-FM exhibits lower performance degradation under external-domain and distribution-shift conditions.
- H1f: XAI outputs identify anatomically and clinically relevant evidence.
- H1g: Uncertainty-aware predictions provide better calibration and more reliable confidence estimates.

3. Problem Statement and Mathematical Formulation

3.1 Multimodal Brain MRI Representation

Let the brain MRI examination of patient i be represented by

$I_i \in R^{M\times H\times W\times D}$,

where:

- M = number of MRI modalities/sequences,
- H = image height,
- W = image width,
- D = depth or number of slices.

For example, M may include T1, T1ce, T2, and FLAIR MRI sequences.

Associated clinical information is represented by

$C_i=\{c_{i1},c_{i2},\ldots,c_{ip}\}$,

where p denotes the number of structured clinical variables.

The complete patient-level multimodal input is therefore

$X_i=\{I_i,C_i\}$.

The objective is to learn a generalizable representation

$Z_i=F_\theta(I_i,C_i)$,

where F_θ represents the NeuroVision-FM encoder with trainable parameters θ.

3.2 MRI Representation ZI

The MRI representation is obtained through two complementary pathways:

$Z_I=F_I(I)=F_{CNN}(I)\oplus F_{ViT}(I)$,

where:

- FCNN extracts local spatial features,
- FViT extracts global contextual features,
- $\oplus$ represents feature integration.

The CNN branch can be represented as

$Z_{CNN}=F_{CNN}(I;\theta_{CNN})$.

The CNN encoder learns hierarchical representations:

$Z_{CNN}(l)=\sigma(W(l)*Z_{CNN}(l-1)+b(l))$,

where:

- W(l) denotes convolution kernels,
- b(l) denotes bias,
- $*$ represents convolution,
- σ represents a nonlinear activation function.

The final CNN representation captures local information including:

- lesion texture,
- tissue boundaries,
- intensity patterns,
- tumor morphology,
- edema characteristics,
- local anatomical structures.

3.3 Transformer-Based Global Representation

MRI features or image patches are converted into tokens:

XP=[x1,x2,…,xN],
where N represents the number of spatial tokens.
Each token is projected into an embedding space:
E=XPWE+Epos,
where:
- □WE is the learnable embedding matrix,
- □Epos represents positional information.

The Transformer uses self-attention:
Q=EWQ,K=EWK,V=EWV.
The attention mechanism is
Attention(Q,K,V)=softmax(dkQKT)V.
Thus,
ZViT=FViT(I;θViT),
where ZViT captures long-range dependencies across spatially distant regions.
The hybrid representation becomes
ZI=ϕ(ZCNN,ZViT),
where ϕ(·) denotes the learned integration operation.

3.4 Clinical Representation ZC
Clinical information may include numerical, categorical, temporal, and textual variables.
For structured information:
C=[c1,c2,…,cp].
After normalization and embedding:
ZC=FC(C;θC).
For a numerical variable:
c~j=σjcj−μj.
Categorical variables can be represented using embeddings:
ej=Emb(cj).
The resulting clinical representation is
ZC=MLP([C~num⊕Ecat]).
This representation enables the model to integrate clinical information with imaging-derived evidence.

3.5 Cross-Modal Attention
Simple concatenation assumes that all modalities contribute independently. NeuroVision-FM instead introduces cross-modal attention.
Let:
ZI∈RN×d
represent imaging features and
ZC∈RL×d
represent clinical features.
Cross-modal attention can be expressed as:
QI=ZIWQ, KC=ZCWK, VC=ZCWV.
The clinical-to-image attention is
AI←C=softmax(dQIKCT).
The clinical-conditioned image representation becomes
Z^I=AI←CVC+ZI.
Similarly, image information can be transferred into the clinical representation:
AC←I=softmax(dQCKIT).
Thus:
Z^C=AC←IVI+ZC.
This bidirectional mechanism allows the model to learn relationships between what is visible in the MRI and what is known clinically about the patient.

3.6 Adaptive Multimodal Feature Fusion
After cross-modal interaction, NeuroVision-FM performs adaptive fusion.
Instead of fixed concatenation:
ZF=ZI⊕ZC,
the model learns modality-specific weights.
Let
αI,αC=softmax(Wf[ZI⊕ZC]+bf).
Then:
ZF=αIZI+αCZC.
with
αI+αC=1.

For multiple MRI sequences, the formulation can be extended as:
ZF=m=1∑MαmZm+αCZC,
subject to
m=1∑Mαm+αC=1.
This mechanism allows NeuroVision-FM to dynamically emphasize the most informative modality for a particular patient.

3.7 Self-Supervised Learning Objective
Self-supervised learning is used to learn representations before supervised adaptation.
A contrastive formulation can be expressed as:
LSSL=−N1i=1∑Nlog∑j=1Nexp(sim(zi,zj)/τ)exp(sim(zi,zi+)/τ),
where:
- zi represents an encoded MRI sample,
- zi+ represents a positive augmentation or corresponding representation,
- sim(·) denotes similarity,
- τ is the temperature parameter.

Alternatively, a masked-image objective can be used:
LMIM=|ΩM|1j∈ΩM∑Ij−I^j22,
where ΩM denotes masked image regions.
For a multimodal model, the self-supervised objective can be expanded:
LSSLmulti=λ1Limage+λ2Lcross−modal.
This enables the model to learn both within-image structure and cross-modal relationships.

3.8 Classification Loss
For a K-class classification task, the standard cross-entropy loss is:
LCLS=−k=1∑Kyklog(y^k),
where:
- yk is the true class indicator,
- y^k is the predicted probability.

For class imbalance, weighted cross-entropy can be used:
LWCE=−k=1∑Kwkyklog(y^k).

3.9 Segmentation Loss
For brain-tumor or lesion segmentation, a combined Dice and cross-entropy loss can be adopted:
LSEG=λDLDice+λCLCE.
Dice loss can be defined as:
LDice=1−∑ipi+∑iyi+ϵ2∑ipiyi+ϵ.
This formulation is useful when pathological regions occupy only a small proportion of the total MRI volume.

3.10 Explainability Objective
Explainability should not be treated simply as generating an attractive heatmap. The framework should evaluate whether explanations are faithful, stable, localized, and clinically meaningful.
Let:
E=G(X,y^)
represent an explanation generated from input X and prediction y^.
A conceptual explanation objective can be expressed as:
LXAI=λFLfaith+λSLstab+λCLclinical.
Here:
- Lfaith measures fidelity of the explanation to the model's prediction,
- Lstab measures explanation stability under small perturbations,
- Lclinical measures agreement with clinically meaningful regions or expert assessment.

For example, if E is an attribution map and R is a reference pathological region, localization overlap can be evaluated using:
IoU(E,R)=|E∪R||E∩R|.
This does not imply that attribution is causal; it provides an empirical measure of anatomical correspondence.

3.11 Uncertainty Estimation
For clinical decision support, the model should estimate uncertainty in addition to prediction probability.
Let:
p(y|X)
represent the predictive distribution.
Predictive entropy can be defined as:
H(Y|X)=−k=1∑Kpklog(pk).
Higher entropy indicates greater predictive uncertainty.
The framework can therefore provide:

Output={y^,p(y|X),U(X),E(X)}.
Thus, every prediction may contain:
- predicted class,
- probability,
- uncertainty,
- visual explanation.

3.12 Overall NeuroVision-FM Optimization Function
The complete model can be trained using a multi-objective optimization function:
$L_{Total}=\lambda_{SSL}L_{SSL}+\lambda_{CLS}L_{CLS}+\lambda_{SEG}L_{SEG}+\lambda_{XAI}L_{XAI}+\lambda_{REG}L_{REG}$
where:
- L_{SSL} = self-supervised representation-learning loss,
- L_{CLS} = classification loss,
- L_{SEG} = segmentation loss,
- L_{XAI} = explainability-related objective,
- L_{REG} = regularization term,
- $\lambda_{SSL},\lambda_{CLS},\lambda_{SEG},\lambda_{XAI},\lambda_{REG}$ = balancing coefficients.

For a classification-only experiment, the formulation can be simplified to:
$L_{Total}=\lambda_{SSL}L_{SSL}+\lambda_{CLS}L_{CLS}+\lambda_{XAI}L_{XAI}+\lambda_{REG}L_{REG}$.
For segmentation:
$L_{Total}=\lambda_{SSL}L_{SSL}+\lambda_{CLS}L_{CLS}+\lambda_{SEG}L_{SEG}+\lambda_{XAI}L_{XAI}+\lambda_{REG}L_{REG}$.

3.13 Complete Mathematical Representation of NeuroVision-FM
The entire architecture can be summarized mathematically as:
$I,C\rightarrow[F_{CNN}(I),F_{ViT}(I),F_C(C)]\rightarrow CrossModalAttention\rightarrow AdaptiveFusion\rightarrow Z_F\rightarrow TaskHead\rightarrow\{\hat{y},U,E\}$
or more formally:
$\hat{y},U,E=H_\psi[F_{Fusion}(F_{CNN}(I),F_{ViT}(I),F_C(C))]$
where H_ψ represents the downstream task-specific prediction head.
The overall training objective is:
$\theta^*=\arg\theta\min L_{Total}$
where θ contains the parameters of the CNN, Transformer, clinical encoder, cross-modal attention, fusion module, and downstream task heads.

3.14 Research Significance of the Mathematical Formulation
The mathematical formulation establishes a clear distinction between the proposed framework and a conventional CNN-based medical classifier.

Component	**Conventional approach**	**NeuroVision-FM**
Learning	Supervised	Self-supervised + supervised adaptation
Image representation	Primarily CNN	CNN + Transformer
Clinical data	Often excluded	Explicit multimodal representation
Fusion	Concatenation/common fusion	Cross-modal attention + adaptive fusion
Transferability	Task-specific	Foundation-model-oriented
Label requirement	High	Reduced through SSL
External validation	Often limited	Core evaluation component
Explainability	Optional	Integrated evaluation
Uncertainty	Often absent	Explicitly modeled
Generalization	Internal performance	Cross-domain/external evaluation

The key scientific proposition is therefore not simply that NeuroVision-FM combines more modules. The central research question is whether self-supervised multimodal representation learning produces a more transferable, robust, interpretable, and clinically useful brain-MRI representation under realistic distributional and annotation constraints.

2. Motivation
- Reduce dependence on expert annotations through self-supervised pretraining.
- Improve cross-domain generalization across scanners, institutions and acquisition protocols.
- Combine complementary imaging and clinical evidence.
- Model local pathology and global anatomical context with CNN and Transformer representations.
- Provide image-region and structured-feature explanations.
- Quantify uncertainty for human-in-the-loop decision support.

- Create a reusable backbone for multiple brain-MRI tasks.

3. Problem Statement

Let I denote a brain MRI examination and C denote associated structured clinical information. A conventional model learns $f(I) \rightarrow y$ for one dataset and task. NeuroVision-FM instead seeks a transferable representation $z=F(I,C)$ such that downstream predictors can be adapted to multiple tasks while remaining robust to distribution shifts. The research problem is to learn a multimodal representation that maximizes diagnostic utility while reducing sensitivity to institution-specific acquisition characteristics and preserving clinically meaningful explanations.

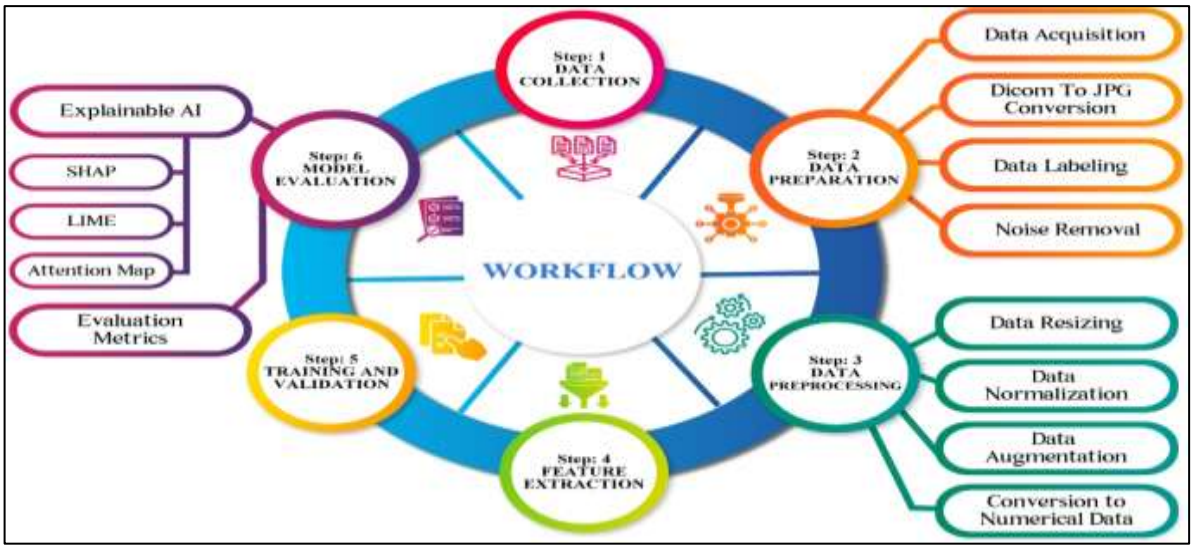
The central research questions are: Can self-supervised pretraining improve label efficiency? Does multimodal fusion improve performance over image-only models? Does combining local CNN features with global Transformer features improve robustness? Can explanations identify clinically meaningful regions and variables? Does the learned representation transfer across datasets and acquisition environments?

4. Main Contributions

- NeuroVision-FM: self-supervised multimodal learning combining CNN local representations, Transformer global context, clinical encoding, cross-modal attention and adaptive fusion.
- Patient-level evaluation to prevent leakage from repeated studies, slices or time points.
- A downstream adaptation strategy for classification and extensible segmentation/localization and risk assessment.
- An explainability layer combining image attribution with structured-feature importance.
- External validation, distribution-shift, missing-modality, calibration and uncertainty evaluation.
- A reproducible experimental design separating illustrative placeholders from measured results.

5. RELATED WORK

CNN-based medical imaging methods established strong baselines for feature extraction and classification, particularly when spatial locality is important. Residual networks and encoder-decoder architectures have been adapted to brain MRI classification and segmentation. Vision Transformers and hybrid CNN-Transformer models provide complementary global contextual modeling through self-attention, although their data and computational requirements can be substantial.



Self-supervised medical representation learning seeks to exploit unlabeled data through contrastive, reconstruction, masked-image, or related objectives. This is relevant because expert MRI annotation is expensive and may contain inter-reader variability. Multimodal foundation-model research additionally aligns images with clinical text, structured information, or other modalities. However, external validation, missing-modality robustness, calibration, and clinically meaningful explainability remain important research challenges.

Explainable AI methods such as Grad-CAM, Integrated Gradients, SHAP and attention visualization can expose attribution patterns. Such attribution should not automatically be interpreted as causal evidence; explanations should be tested for stability, faithfulness, anatomical plausibility and clinical usefulness.

Table 1. Comparison of Existing Methods

Method	Strengths	Limitations	Generalization	Explainability
CNN	Strong local features	Limited global context	Moderate	Grad-CAM
Vision Transformer	Global attention	Data/compute intensive	Moderate–High with pretraining	Attention/attribution
CNN–Transformer hybrid	Local + global representation	Higher complexity	Potentially high	Multiple attribution paths

Self-supervised MRI	Uses unlabeled data	Objective/domain dependent	Potentially high	Post-hoc XAI
Image-only	Simple deployment	Ignores clinical context	Often limited	Image attribution
Multimodal fusion	Complementary evidence	Missing modalities/alignment	Potentially high	Feature + image XAI
NeuroVision-FM	SSL + hybrid + multimodal + XAI	Requires rigorous validation	Designed for cross-domain tests	Integrated evaluation

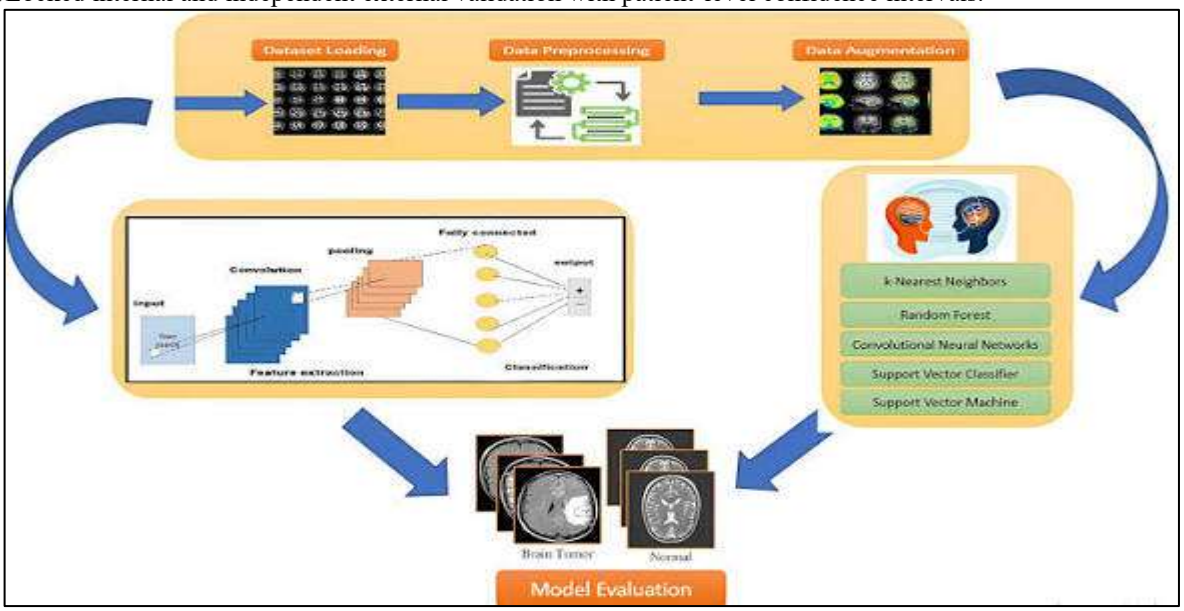
7. Research Gap

- Task-specific rather than reusable representations.
- Dependence on expensive expert labels.
- Limited clinical-context integration.
- Risk of leakage under image-level splitting.
- Insufficient external validation in many studies.
- Explainability often lacks quantitative or expert validation.
- Calibration and uncertainty are not consistently integrated.
- Missing-modality and acquisition-shift robustness need systematic analysis.

8. Proposed Framework for Machine Learning Based Explainable Brain MRI Analysis

NeuroVision-FM contains seven principal stages: data governance and patient indexing; modality-specific preprocessing; self-supervised visual pretraining; hybrid CNN-Transformer representation learning; clinical feature encoding and cross-modal attention; adaptive multimodal fusion; and task-specific prediction with explainability and uncertainty estimation.

- Data governance: de-identification, patient indexing, duplicate resolution and cohort definition.
- MRI preprocessing: orientation standardization, resampling, intensity normalization, quality control and medically justified augmentation.
- Self-supervised pretraining on unlabeled MRI using a documented objective such as masked-image modeling or contrastive learning.
- Hybrid visual encoder: CNN local features plus Vision Transformer global features.
- Clinical encoder: structured variables transformed into a latent representation with leakage-safe missing-value handling.
- Cross-modal attention and adaptive fusion to learn interactions and dynamic modality weighting.
- Downstream heads for classification, segmentation/localization or risk prediction with uncertainty and XAI.
- Locked internal and independent external validation with patient-level confidence intervals.



9. Algorithm 1: Explainable Machine Learning Brain MRI Assessment Model

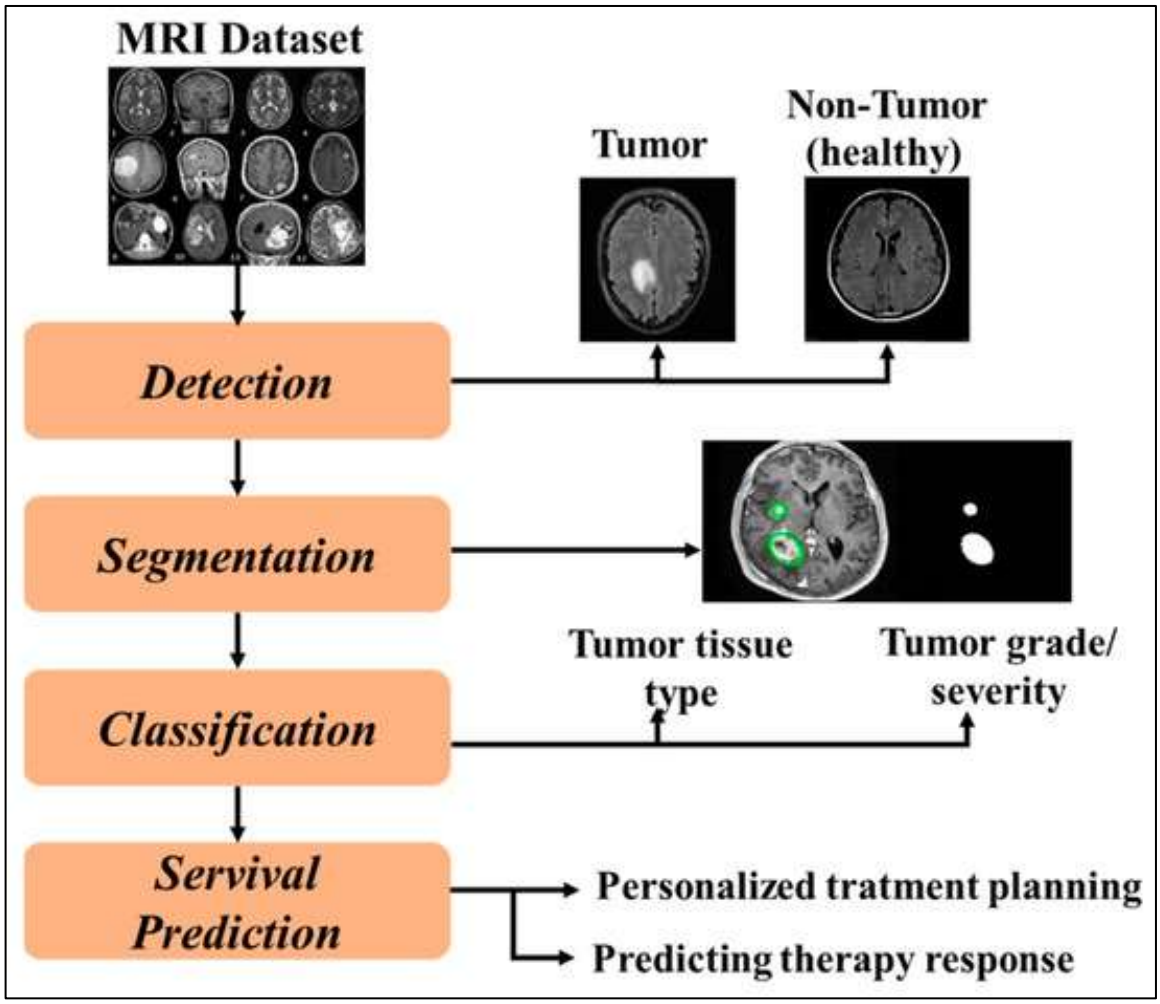
Input: MRI examinations I, clinical variables C, optional text T, labels Y, patient IDs P. Output: prediction $\hat{y}$, probability p, uncertainty u, explanation E.

1. Group all observations by patient and create patient-level partitions.
2. Perform MRI quality control, spatial/intensity normalization and medically valid augmentation.
3. Pretrain the visual encoder using the selected self-supervised objective.
4. Extract local CNN representation and global ViT representation.
5. Encode clinical variables and optional clinical text.
6. Apply cross-modal attention and adaptive fusion.

7. Optimize the downstream objective using training data and validation-only model selection.
8. Calibrate probabilities where appropriate.
9. Generate prediction and uncertainty.
10. Generate image and feature explanations using suitable XAI methods.
11. Evaluate locked test and external cohorts using sensitivity, specificity, F1, ROC-AUC, PR-AUC, calibration, robustness and confidence intervals.
12. Perform ablation, subgroup and failure-mode analyses.

10. Evaluation Setup — Datasets

Candidate public resources include BraTS for brain-tumor segmentation and related MRI research, IXI for multi-site structural MRI representation learning, and ADNI for Alzheimer's/dementia-related multimodal research subject to access conditions. The exact dataset selection must match the target task and current licensing/data-use requirements. If multiple datasets are used, development and external-validation cohorts should be clearly separated and checked for patient overlap.



Patient-level splitting is mandatory when multiple slices, sequences, studies or time points belong to one subject. Preprocessing statistics, feature selection and hyperparameter tuning must be derived without access to the locked test set.

Table 2. Simulation Environment

Component	Recommended configuration	Final paper should report
OS	Ubuntu 22.04 LTS / Windows 11	Exact OS/version
Language	Python 3.11	Exact version
DL framework	PyTorch 2.x	Exact version
Medical imaging	MONAI / NiBabel / SimpleITK	Versions
GPU	NVIDIA CUDA GPU	Model + VRAM
CUDA	Compatible CUDA 12.x	Exact version
CPU/RAM	Multi-core / ≥32 GB	Models/capacity
Reproducibility	Fixed seeds	Seeds + determinism
Statistics	Patient-level bootstrap	Bootstrap count + CI method

12. Performance Evaluation

Primary classification measures should include accuracy, precision, sensitivity/recall, specificity, F1-score, ROC-AUC and PR-AUC. For imbalanced biomedical data, class-wise sensitivity and PR-AUC should be emphasized. Calibration

should be evaluated with reliability diagrams and expected calibration error. If segmentation is performed, Dice, IoU/Jaccard, HD95 and sensitivity should be reported. Major metrics should include 95% confidence intervals, preferably from patient-level bootstrap resampling. External validation and robustness to missing modalities and distribution shift should be central analyses.

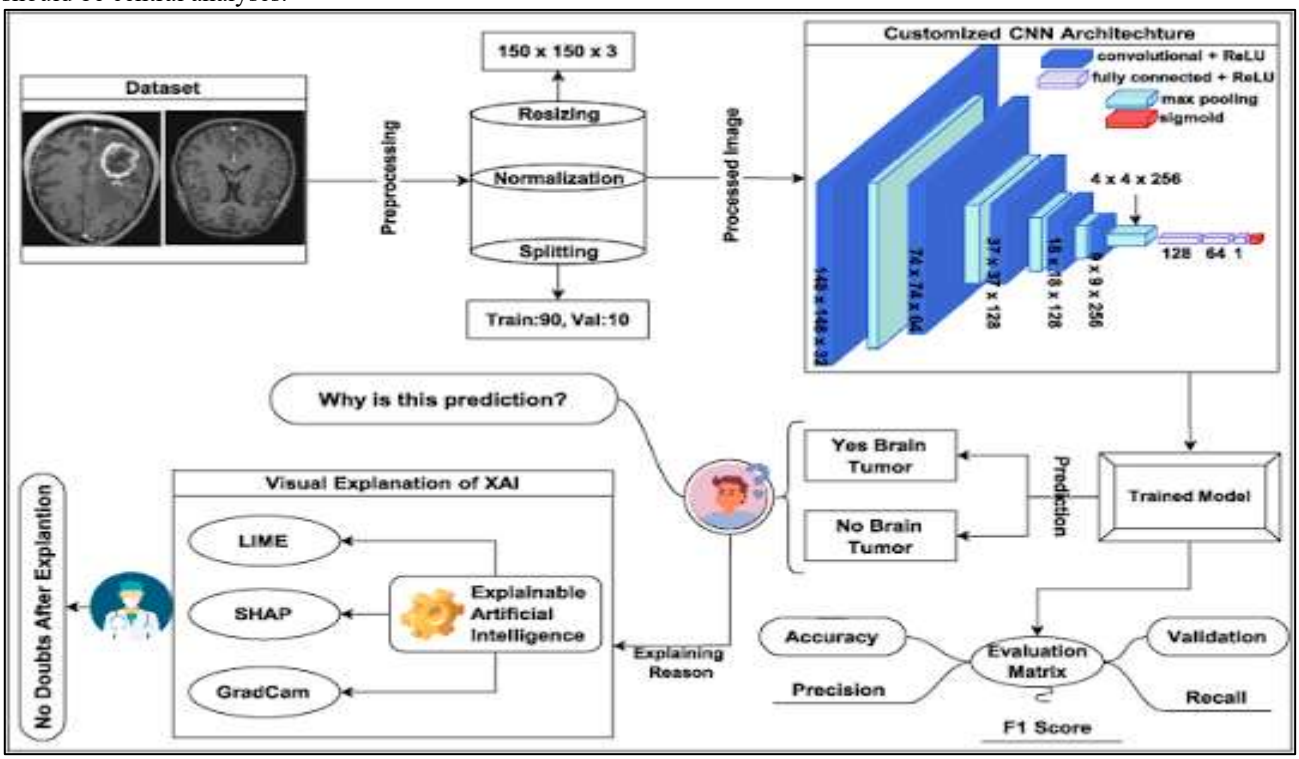
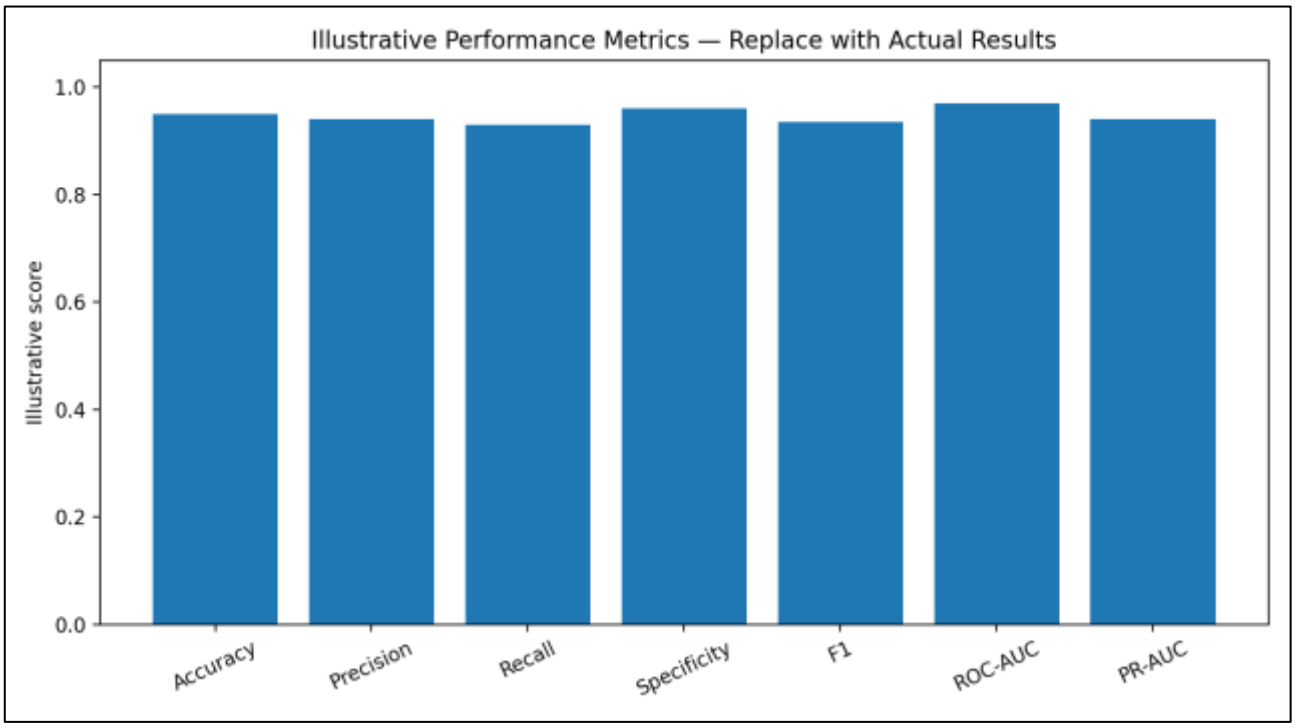


Figure 1. Analysis of Performance Metrics



This figure is illustration values with experimental results.

14. Interpretations and Outcomes

Results should be interpreted across multiple dimensions rather than through accuracy alone. An increase in accuracy without improved sensitivity or PR-AUC may not be meaningful for an imbalanced diagnostic task. A higher ROC-AUC with poor calibration may indicate good ranking but unreliable probability estimates. A convincing study should show consistent performance, external generalization, calibration, robustness, and clinically plausible explanations.

Ablation studies should establish whether gains originate from self-supervised pretraining, Transformer context, clinical fusion, cross-modal attention or adaptive fusion rather than merely from increased parameter count. Failure cases should be inspected for artifacts, acquisition-specific cues, anatomical mislocalization and clinically ambiguous cases.

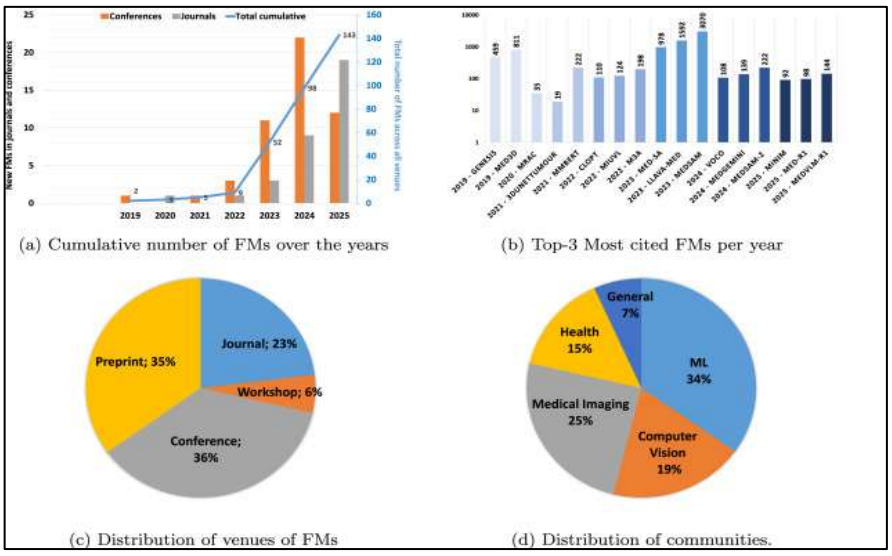
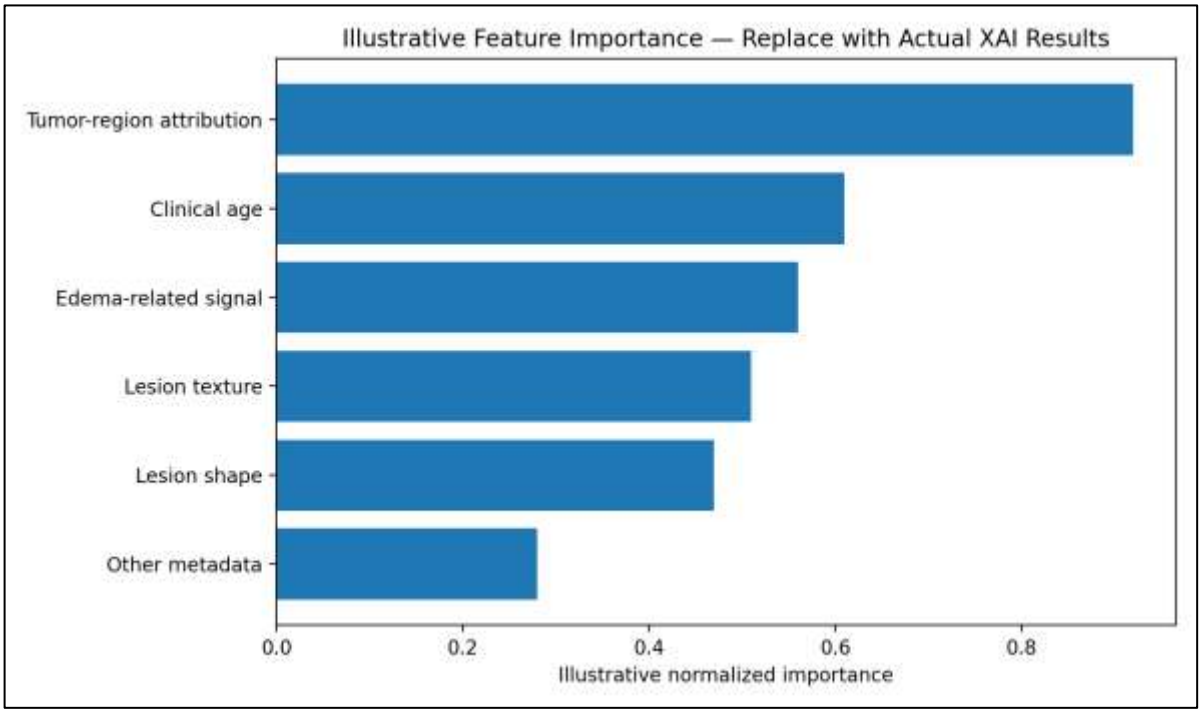


Figure 2. Feature Importance Analysis



This feature-importance chart is illustration of actual SHAP/Integrated-Gradients/Grad-CAM-derived values must be generated from the trained model and real test cohort.

Table 3. Clinical Interpretability

Evidence	XAI method	Interpretation target	Validation
MRI image	Grad-CAM/Grad-CAM++	Lesion/anatomical region	Expert region agreement
Transformer	Attention/attribution	Global contextual evidence	Stability analysis
Clinical variables	SHAP/Integrated Gradients	Feature contribution	Clinical plausibility
Multimodal fusion	Cross-modal attribution	Image–clinical interaction	Ablation + expert review
Probability	Calibration + uncertainty	Confidence of prediction	Reliability/ECE
Failures	Counterfactual/error analysis	Failure mechanism	Expert review where feasible

17. Statistical Analysis

Report 95% confidence intervals for major metrics. Patient-level bootstrap is recommended when subjects contribute multiple observations. Model comparisons should use paired patient-level evaluation where possible. Appropriate analyses may include DeLong's test for correlated ROC-AUCs, McNemar's test for paired classification outcomes, bootstrap comparisons, or permutation testing. Report effect sizes and confidence intervals and control multiplicity where necessary.

18. CONCLUSION

NeuroVision-FM presents a research framework for generalizable brain MRI analysis based on self-supervised multimodal representation learning. The framework combines CNN local feature extraction, Transformer global contextual modeling, structured clinical encoding, cross-modal attention, adaptive fusion, uncertainty estimation and explainable AI. Its central premise is that reusable representations learned from diverse unlabeled and multimodal medical information can reduce annotation dependence and improve transferability across downstream brain-MRI tasks.

The architecture should be considered a proposed research framework until validated through real experiments. A rigorous study should use patient-level separation, external validation, calibration, robustness testing, ablation studies, confidence intervals and clinically meaningful explainability analysis. Only after reproducible evidence is obtained should claims about foundation-model capability or clinical deployment be made.

19. FUTURE WORK

- Large-scale multi-institutional self-supervised pretraining across diverse MRI sequences and scanners.
- Integration of radiology reports and medical-language modalities.
- Federated/privacy-preserving pretraining across institutions.
- Continual learning and domain adaptation.
- Advanced uncertainty estimation and selective prediction.
- Prospective multi-reader studies of explanation usefulness.
- Fairness and subgroup analysis.
- Parameter-efficient adaptation and model compression.
- Geographically and institutionally distinct external validation.
- Prospective clinical utility and health-economic evaluation.

REFERENCES

1. Dr. Kommu Naveen & Dr. RMS Parvathi, Scientific Reports-Springer Nature, AI Powered Multi Feature Fusion Framework for Retrieving Images Using Color, Texture and Shape Descriptors. || ISSN 2984-8687 || © November 2025. SCI-Q1 (2025).
2. Litjens, G., et al., “A survey on deep learning in medical image analysis,” Medical Image Analysis, 2017.
3. Ronneberger, O., Fischer, P., and Brox, T., “U-Net: Convolutional Networks for Biomedical Image Segmentation,” MICCAI, 2015.
4. He, K., Zhang, X., Ren, S., and Sun, J., “Deep Residual Learning for Image Recognition,” CVPR, 2016.
5. Dosovitskiy, A., et al., “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” ICLR, 2021.
6. Hatamizadeh, A., et al., “UNETR: Transformers for 3D Medical Image Segmentation,” WACV, 2022.
7. Isensee, F., et al., “nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation,” Nature Methods, 2021.
8. Menze, B. H., et al., “The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS),” IEEE Transactions on Medical Imaging, 2015.
9. Selvaraju, R. R., et al., “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization,” ICCV, 2017.
10. Lundberg, S. M. and Lee, S.-I., “A Unified Approach to Interpreting Model Predictions,” NeurIPS, 2017.
11. Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q., “On Calibration of Modern Neural Networks,” ICML, 2017.
12. Azizi, S., et al., “Big Self-Supervised Models Advance Medical Image Classification,” ICCV, 2021.