CINEFORUM
ISSN: 0009-7039
Vol. 66. No. S3, 2026 (Special Issue 3)

Multimodal Medical Foundation Models For Biomedical Image Analysis: An Artificial Intelligence And Deep Learning Framework

Dr. Kommu Naveen[1], Dr. G. Karthik Reddy[2], Dr. Gurmeet Singh[3], Dr. Pujari. Lakshmi Devi[4], Dr. Dharavath Srinivas[5], Chiranjeevi Thokala[6], Shireesha Jalda[7], Dr. Battula Balnarsaiah[8]

[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]Associate Professor, Department of Electronics and Communication Engineering MLR Institute of Technology, Laxman Reddy Avenue, Dundigal Police Station Road, Quthbullapur Mandal, Medchal-Malkajgiri district, Hyderabad, Telangana, 500043, India.gkarthik.ouece@gmail.com
[3]Associate Professor, Department of Electronics and Communication Engineering, Guru Tegh Bahadur institute of technology, Delhi - 110064, India. drgurmeetsingh@gtbit.ac.in
[4]Professor, Department of Electronics and Communication Engineering, St.Peter's Engineering college. Maisummaguda, Opposite to Forest Academy, Kompally, Hyderabad. Telangana, India-500100. lakshmipujaridhar03@gmail.com
[5]Assistant Professor, Department of Electronics and Communication Engineering, Guru Nanak Institute of Technology, Khanapur Village, Ibrahimpatnam, Ranga Reddy District, Hyderabad, Telangana – 501506, India. sreenuchandra91@gmail.com
[6]Assistant Professor, Department of Electronics and Communication Engineering, VNR VJIET, Bachupally, Hyderabad -500118, India. chiranjeevi_t@vnrvjiet.in
[7]Assistant Professor, Department of Electronics and Communication Engineering, TKR College Of Engineering And Technology, Survey No -8/A Medbowli, Meerpet, Telangana 500097, India. shirishajalda@gmail.com
[8]Associate Professor, Department of AIML, Nalla Malla Reddy Engineering College, Divyanagar, Narapally, Hyderabad-Telanagana State, India – 500088, battulabalu@gmail.com

Abstract
The rapid development of artificial intelligence (AI), deep learning (DL), and multimodal representation learning is reshaping biomedical image analysis from task-specific classifiers toward general-purpose medical foundation models. Conventional convolutional neural networks (CNNs) can learn effective local visual representations, while Vision Transformers (ViTs) and related attention architectures provide global contextual modeling; however, isolated models often remain constrained by limited annotations, domain shift, heterogeneous modalities, incomplete clinical context, and weak interpretability. This paper proposes a multimodal medical foundation model framework, termed MedFusion-FM, for biomedical image analysis that integrates self-supervised visual pretraining, CNN-based local feature extraction, Transformer-based global representation learning, cross-modal attention, adaptive feature fusion, and explainable AI. The framework is designed to jointly exploit medical images with structured clinical metadata and, where available, electronic health records and medical text. The workflow includes patient-level data partitioning, modality harmonization, quality control, augmentation, self-supervised pretraining, multimodal alignment, task-specific adaptation, uncertainty estimation, and explainability. Evaluation is specified using accuracy, precision, recall, F1-score, specificity, ROC-AUC, PR-AUC, calibration, computational cost, and robustness under distribution shift. Illustrative figures and benchmark values are included only to demonstrate manuscript presentation and must be replaced by actual experimental results before submission. The proposed research direction aims to establish a reusable and clinically responsible foundation-model paradigm for diagnosis, segmentation, prognosis, and decision support across biomedical imaging applications. The performance of the proposed MedFusion-FM framework should be evaluated using a comprehensive set of discrimination, calibration, segmentation, robustness, explainability, and computational metrics rather than relying exclusively on classification accuracy. This multidimensional evaluation is particularly important for biomedical datasets, where class imbalance, patient-level correlation, acquisition heterogeneity, incomplete modalities, and domain shift can substantially influence the apparent performance of an artificial intelligence model. The F1-score provides a balanced summary of precision and recall and is especially useful for imbalanced datasets. For multi-class medical classification, macro-averaged and weighted-averaged F1-scores should be reported together with class-wise results. ROC-AUC should be used to assess discrimination across classification thresholds, while PR-AUC should receive particular attention when the positive class is rare. In such

circumstances, PR-AUC may provide a more informative assessment of minority-class detection than accuracy or ROC-AUC alone.

Keywords: Multimodal Medical Foundation Models; Biomedical Image Analysis; Artificial Intelligence; Deep Learning; CNN; Vision Transformer; Self-Supervised Learning; Cross-Modal Attention; Explainable AI; Medical Imaging; Clinical Decision Support.

Introduction

Biomedical imaging is central to diagnosis, staging, treatment planning, monitoring, and prognosis. MRI, CT, radiography, ultrasound, mammography, dermoscopy, ophthalmic imaging, and digital pathology contain complex spatial and morphological patterns that are difficult to quantify consistently by manual inspection alone. Deep learning has demonstrated substantial capability in extracting diagnostic representations from these data, particularly through convolutional neural networks. Nevertheless, many published systems are designed for one dataset, one modality, and one narrowly defined prediction task. Such specialization can produce high internal test performance without guaranteeing generalization across hospitals, scanners, populations, acquisition protocols, or clinical workflows.

Medical foundation models offer a broader paradigm. Rather than training a separate model from scratch for every task, a foundation model can learn transferable representations from large and heterogeneous datasets and subsequently be adapted to classification, segmentation, detection, retrieval, report generation, or prognosis. Multimodal foundation models extend this concept by aligning visual information with clinical text, structured patient information, laboratory observations, and other contextual sources. This paper therefore frames biomedical image analysis as a multimodal representation-learning problem rather than a simple image-classification problem.

2. Motivation

- Reduce dependence on large manually annotated datasets through self-supervised and transfer learning.
- Improve generalization across institutions, imaging devices, acquisition protocols, and patient populations.
- Integrate image evidence with clinical context instead of relying exclusively on pixels.
- Provide interpretable and uncertainty-aware outputs suitable for clinician review.
- Create reusable representations that support multiple downstream biomedical tasks.
- Investigate computationally efficient adaptation for practical clinical deployment.

2. Background and Related Research

2.1 Deep Learning for Biomedical Image Analysis

CNNs introduced strong inductive biases for spatial feature extraction and remain effective for many medical imaging tasks. Encoder-decoder architectures have become particularly influential for segmentation, while residual and densely connected architectures improve optimization and representation depth. However, convolutional receptive fields and local operations can make explicit modeling of very long-range relationships more difficult than attention-based approaches.

2.2 Vision Transformers

Vision Transformers tokenize images into patches or learned visual tokens and apply self-attention to model relationships between distant regions. Their success is strongly associated with large-scale pretraining and suitable regularization. Medical applications must nevertheless consider the relatively small size of many labeled datasets and the high computational cost of three-dimensional imaging.

2.3 Self-Supervised Medical Representation Learning

Self-supervised learning reduces dependence on manual annotations by constructing training objectives from the data itself. Masked image modeling, contrastive learning, multi-view consistency, and image-text alignment can encourage representations that transfer to downstream tasks. In biomedical applications, the pretraining corpus must be carefully governed to avoid patient leakage and hidden overlap with evaluation data.

2.4 Multimodal Medical Foundation Models

Multimodal modeling seeks a shared or coordinated representation across modalities. A medical image can be paired with clinical variables, radiology reports, pathology descriptions, laboratory results, genomic information, or longitudinal records. Cross-modal learning can improve contextual interpretation, but modality alignment is challenging because observations may be asynchronous, incomplete, noisy, or weakly paired.

3. Problem Statement

The central research problem is to develop a scalable multimodal medical foundation-model architecture that can learn robust representations from heterogeneous biomedical images and complementary clinical information while maintaining diagnostic discrimination, cross-domain generalization, interpretability, uncertainty awareness, and

computational feasibility. The model must avoid patient-level leakage, accommodate missing or noisy modalities, and provide statistically defensible evaluation rather than relying on accuracy alone.

4. Main Contributions

- MedFusion-FM: a proposed multimodal medical foundation-model framework for biomedical image analysis.
- Hybrid CNN–Transformer visual encoder for complementary local and global representation learning.
- Cross-modal attention mechanism for image–clinical information alignment.
- Adaptive fusion and modality-masking strategy for incomplete clinical data.
- Self-supervised pretraining and parameter-efficient downstream adaptation.
- Explainability and uncertainty layer intended to support responsible clinical interpretation.
- Evaluation protocol covering discrimination, calibration, robustness, efficiency, and ablation analysis.

5. Comparison of Existing Methods

Method	Strengths	Limitations	Multimodal	Generalization	Typical Use
CNN	Strong local texture/morphology learning; efficient	Limited global context; task-specific	Low	Moderate	Classification/segmentation
Vision Transformer	Global attention; scalable representation learning	Data/computation intensive	Possible	High with pretraining	Image classification
CNN–Transformer Hybrid	Local + global features	More parameters; fusion complexity	Possible	High	Medical image diagnosis
Self-Supervised Models	Uses unlabeled data; label efficient	Pretext-task sensitivity	Possible	High	Pretraining
Multimodal Foundation Model	Joint image/context representation; reusable	Large data, compute and alignment needs	Yes	Potentially very high	Diagnosis/retrieval/reporting
Proposed MedFusion-FM	Hybrid, multimodal, self-supervised, XAI, adaptable	Requires rigorous external validation	Yes	Target: high	Multi-task biomedical analysis

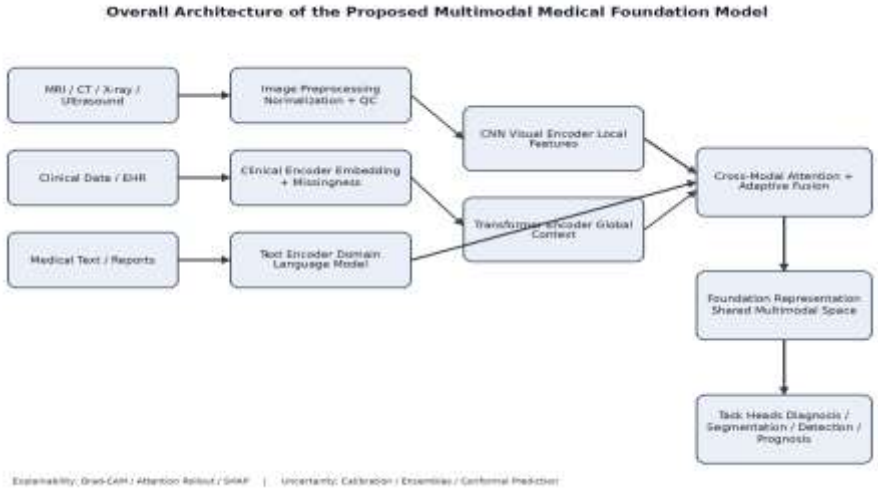
6. Research Gap

The literature shows a transition from conventional CNN classifiers to Transformer and foundation-model approaches, but several gaps remain. Many biomedical systems still evaluate a single modality and a single task. Models can exhibit substantial performance degradation under domain shift. Multimodal studies may assume that every modality is always available, although this is unrealistic in clinical practice. High predictive performance does not automatically establish explanation faithfulness or calibration. Many studies also lack patient-level leakage controls, external validation, confidence intervals, or systematic ablation. The proposed framework addresses these gaps through multimodal alignment, modality masking, self-supervised pretraining, external-domain testing, explainability, uncertainty analysis, and reproducible evaluation.

7. Proposed Framework for Artificial Intelligence–Deep Learning

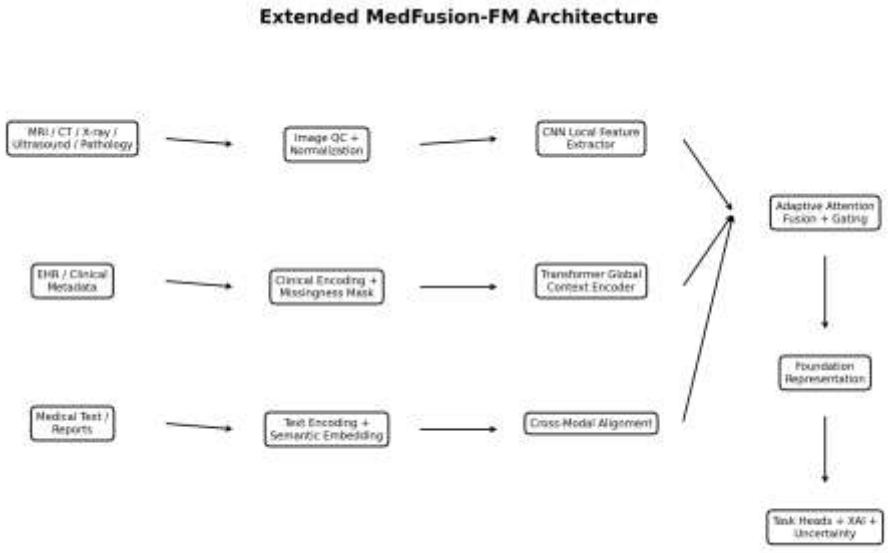
The proposed MedFusion-FM consists of five major stages: heterogeneous biomedical data acquisition and

governance; modality-specific preprocessing and quality control; self-supervised visual representation learning using a hybrid CNN–Transformer encoder; cross-modal attention and adaptive feature fusion; and task-specific prediction with explainability and uncertainty estimation. The framework is intended as a research architecture rather than a clinically validated medical device.



7.1 Architecture

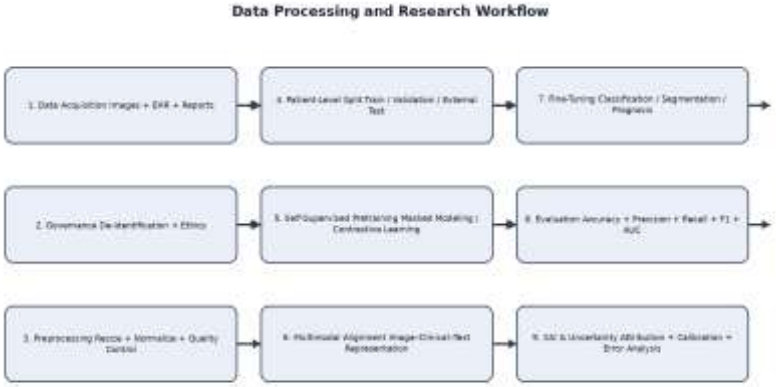
- Visual encoder: convolutional stem extracts fine-grained edges, textures, and local anatomical patterns.
- Transformer encoder: multi-head self-attention models long-range spatial dependencies.
- Clinical encoder: structured variables are normalized and embedded; text can be encoded by a domain-adapted language encoder.
- Cross-modal attention: image tokens query clinical/context tokens and vice versa.
- Adaptive fusion: learned gating weights control the contribution of each modality.
- Prediction heads: separate heads can support classification, segmentation, detection, prognosis, or retrieval.
- XAI and uncertainty: attribution maps, attention analysis, confidence calibration, and uncertainty estimates accompany predictions.



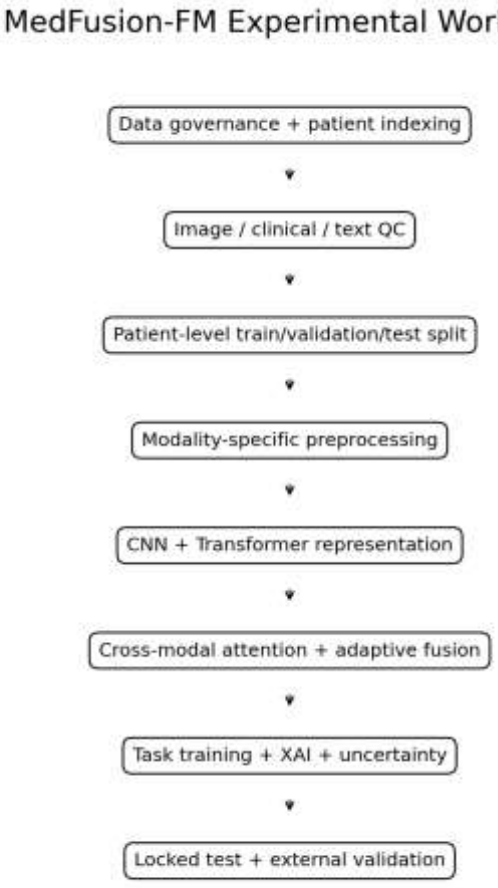
8. Data Processing and Workflow

A rigorous workflow should begin with ethics, data governance, de-identification, and patient-level indexing.

Images should be converted to a consistent representation, checked for corrupted studies, harmonized for orientation and resolution, and normalized according to modality-specific practice. Clinical variables should be standardized, categorical values encoded, missingness explicitly represented, and free text de-identified. Splitting must occur at the patient level before augmentation or model fitting to prevent information leakage. Training data may then undergo controlled augmentation, while validation and test data should represent the intended deployment distribution.



- Data acquisition and governance → de-identification → patient-level linkage → image quality assessment → preprocessing → clinical metadata cleaning → patient-level partitioning → self-supervised pretraining → task adaptation → internal/external evaluation → explainability and uncertainty analysis.



9. Algorithms and Mathematical Formulations

9.1 CNN Local Feature Extraction

For an input image X, convolutional features can be represented as $F_{CNN} = Conv(X; \theta_C)$, where θ_C denotes learnable convolutional parameters. Multi-scale convolutional blocks can preserve fine and intermediate anatomical structures.

9.2 Transformer Global Representation

Given token embeddings Z, self-attention is computed as $Attention(Q,K,V) = softmax(QK^T/\sqrt{d})V$, where Q, K, and V are query, key, and value matrices and d is the key dimension. Multi-head attention allows complementary relational patterns to be learned.

9.3 Adaptive Cross-Modal Fusion

Let V represent visual embeddings and C represent clinical embeddings. Cross-modal attention produces $A_{VC} = softmax(Q_V K_C^T/\sqrt{d})V_C$. A learnable gate g can then combine modalities as $F = g \cdot V + (1-g) \cdot C$, with g constrained to [0,1].

9.4 Multi-Task Objective

A practical objective can combine classification, segmentation, alignment, and regularization terms: $L = \lambda_{cls} L_{cls} + \lambda_{seg} L_{seg} + \lambda_{align} L_{align} + \lambda_{reg} L_{reg}$. The coefficients λ should be selected using validation data rather than test performance.

9.5 Self-Supervised Learning

Self-supervised pretraining can use masked image modeling, contrastive learning, multi-view consistency, or image–text alignment. The purpose is to learn transferable representations before downstream labels are introduced. Pretraining and downstream evaluation should be strictly separated to avoid test-data contamination.

9.6 Pseudocode

Algorithm 1 — Training:

Input: image I, clinical C, optional text T, labels Y.
1. Group records by patient. - For an input image X, convolutional features can be represented as $F_{CNN} = Conv(X; \theta_C)$,
2. Create patient-level train/validation/test partitions.
3. Preprocess each modality.
4. Generate modality embeddings Z, self-attention is computed as $Attention(Q,K,V) = softmax(QK^T/\sqrt{d})V$
5. Apply cross-modal attention - V represent visual embeddings and C represent clinical embeddings. Cross-modal attention produces $A_{VC} = softmax(Q_V K_C^T/\sqrt{d})V_C$
6. Compute adaptive fusion.
7. Optimize classification/segmentation/multitask loss - $L = \lambda_{cls} L_{cls} + \lambda_{seg} L_{seg} + \lambda_{align} L_{align} + \lambda_{reg} L_{reg}$.
8. Select using validation data.
9: $F* \leftarrow \{Clinical, Troponin, Systolic\ BP, Smk\ Status\}$
10: Initialize H attention heads
11: for $h = 1$ to H do
12: $Qh \leftarrow XWQh, Kh \leftarrow XWKh, Vh \leftarrow XWVh$
13: $\alpha h \leftarrow Softmax((QhKhT)/\sqrt{dk})$
14: $Zh \leftarrow \alpha h \times Vh$
15. Lock model and evaluate on test data.
16. Repeat with external data where available.
17. Report confidence intervals, ablations, robustness and computational cost.

9.7 Starting Hyper Parameters

Parameter	Starting value	Candidate range	Report
Batch size	16	8–64	Yes
Learning rate	1e-4	1e-5–1e-3	Yes
Epochs	20	10–100	Yes
Image size	224×224	Task-dependent	Yes
Optimizer	AdamW	Adam/AdamW	Yes
Fusion dimension	256	128–768	Yes
Transformer heads	8	4–16	Yes

Random seed	42	Predefined seeds	Yes

10. Performance Evaluation

The primary classification measures should include accuracy, precision, recall (sensitivity), specificity, F1-score, ROC-AUC, and PR-AUC. For imbalanced biomedical datasets, PR-AUC and class-wise sensitivity can be more informative than accuracy alone. Calibration should be assessed using reliability diagrams and expected calibration error. For segmentation, Dice coefficient, Jaccard/IoU, Hausdorff distance, and sensitivity should be reported. Confidence intervals, preferably obtained through patient-level bootstrap resampling, should accompany major metrics.

The performance of the proposed MedFusion-FM framework should be evaluated using a comprehensive set of discrimination, calibration, segmentation, robustness, explainability, and computational metrics rather than relying exclusively on classification accuracy. This multidimensional evaluation is particularly important for biomedical datasets, where class imbalance, patient-level correlation, acquisition heterogeneity, incomplete modalities, and domain shift can substantially influence the apparent performance of an artificial intelligence model.

For classification tasks, the primary evaluation measures should include accuracy, precision, recall (sensitivity), specificity, F1-score, receiver operating characteristic area under the curve (ROC-AUC), and precision–recall area under the curve (PR-AUC). Accuracy represents the proportion of correctly classified observations; however, it can be misleading when the class distribution is highly imbalanced. Therefore, precision and recall should be reported to characterize false-positive and false-negative behavior. Sensitivity is particularly important in diagnostic and screening applications because failure to identify a disease-positive patient may have substantial clinical consequences. Specificity complements sensitivity by quantifying the ability of the model to correctly identify negative cases.

The F1-score provides a balanced summary of precision and recall and is especially useful for imbalanced datasets. For multi-class medical classification, macro-averaged and weighted-averaged F1-scores should be reported together with class-wise results. ROC-AUC should be used to assess discrimination across classification thresholds, while PR-AUC should receive particular attention when the positive class is rare. In such circumstances, PR-AUC may provide a more informative assessment of minority-class detection than accuracy or ROC-AUC alone.

The confusion matrix should be reported to provide a detailed representation of true-positive, true-negative, false-positive, and false-negative predictions. For multi-class applications, the confusion matrix should be extended to all diagnostic categories, allowing systematic analysis of disease-specific misclassification patterns. Such analysis is important for identifying classes that remain difficult to distinguish and for determining whether errors are associated with disease similarity, insufficient training samples, image quality, or domain variation.

Probability calibration should also be evaluated because a model with strong discrimination may nevertheless produce unreliable confidence estimates. Reliability diagrams should compare predicted probabilities with observed event frequencies across probability intervals. Expected calibration error (ECE) can be reported as a quantitative summary of calibration deviation. The Brier score may additionally be used to assess the quality of probabilistic predictions. Lower ECE and Brier scores indicate better probabilistic reliability. Calibration is particularly important for clinical decision-support applications because clinicians may use predicted probabilities to determine the need for further investigation or expert review.

If the proposed framework is extended to biomedical image segmentation, Dice similarity coefficient, Jaccard index or intersection-over-union (IoU), Hausdorff distance, and sensitivity should be reported. Dice and IoU measure spatial overlap between predicted and expert-annotated regions, whereas Hausdorff distance provides information about boundary discrepancies. The 95th-percentile Hausdorff distance (HD95) may be preferred to the maximum Hausdorff distance because it reduces the influence of isolated boundary outliers. Sensitivity and specificity should additionally be reported to characterize voxel-level or pixel-level detection behavior.

Statistical uncertainty should accompany the major performance estimates. Instead of reporting only point estimates, the study should provide 95% confidence intervals for accuracy, precision, recall, specificity, F1-score, ROC-AUC, PR-AUC, Dice, and IoU where appropriate. Patient-level bootstrap resampling is recommended when multiple images, examinations, or observations are associated with the same patient. In patient-level bootstrapping, patients rather than individual images are sampled with replacement, and the performance metric is recalculated over a large number of bootstrap replications. The resulting empirical distribution can then be used to estimate confidence intervals. This approach reduces the risk of artificially narrow confidence intervals caused by treating correlated observations from the same patient as independent.

The evaluation should further include external validation whenever an independent dataset is available. The external cohort should ideally differ from the development dataset in institution, scanner, acquisition protocol, patient population, or other relevant characteristics. Performance consistency between internal and external

datasets provides stronger evidence of model generalization than performance on a single randomly divided dataset. A substantial reduction in external performance should be investigated through domain-shift and failure-mode analysis rather than concealed by reporting only the best-performing internal result.

Robustness analysis should evaluate MedFusion-FM under image noise, reduced resolution, acquisition variability, intensity perturbation, missing clinical variables, and missing imaging modalities. Since multimodal medical data are frequently incomplete in practical clinical environments, missing-modality experiments are particularly relevant. The model should be evaluated using complete multimodal input as well as systematically masked configurations to determine whether the adaptive fusion mechanism can maintain useful diagnostic performance when one or more modalities are unavailable.

Explainability should be evaluated alongside predictive performance. For image-based predictions, Grad-CAM, Grad-CAM++, attention visualization, or related attribution methods can be used to identify image regions contributing to model predictions. For structured clinical variables, SHAP or integrated-gradient-based approaches can provide complementary feature-level explanations. Where feasible, explainability should be assessed for localization quality, stability, faithfulness, and clinical plausibility rather than assuming that visually appealing heatmaps automatically represent causal evidence.

Finally, computational performance should be reported to establish the practical feasibility of the proposed foundation-model framework. Relevant measures include the number of trainable parameters, computational complexity, GPU memory consumption, training time, inference latency, throughput, and model storage requirements. These measurements are important because multimodal foundation models can provide improved representation capacity at the cost of increased computational requirements. Reporting both predictive performance and computational efficiency enables a more balanced comparison with CNN-only, Transformer-only, and conventional multimodal fusion approaches.

Accordingly, the final evaluation of MedFusion-FM should combine discrimination, calibration, segmentation quality, statistical confidence, external generalization, robustness, explainability, and computational efficiency. Such a comprehensive evaluation strategy provides a more rigorous assessment of whether the proposed framework represents a genuine improvement over existing biomedical image-analysis methods and whether its predictions are sufficiently reliable for further clinical investigation.

For **MedFusion-FM**, performance evaluation should go beyond reporting a single accuracy value. Since biomedical datasets are often **imbalanced, heterogeneous, multimodal, and clinically sensitive**, the evaluation should measure discrimination, class-wise performance, calibration, segmentation quality, robustness, and statistical uncertainty.

1. Classification Performance Metrics

Let:

- **TP** = True Positives
- **TN** = True Negatives
- **FP** = False Positives
- **FN** = False Negatives

The confusion matrix is the fundamental basis for most classification metrics.

Actual / Predicted	**Positive**	**Negative**
Positive	TP	FN
Negative	FP	TN

1.1 Accuracy

Accuracy measures the proportion of all correctly classified samples.

Accuracy is easy to interpret, but it can be **misleading for imbalanced medical datasets**.

For example, suppose 95% of patients are negative and only 5% are positive. A model that predicts every patient as negative can achieve 95% accuracy while completely failing to detect the disease.

Therefore, accuracy should **not be used as the sole primary metric**.

1.2 Precision

Precision represents the proportion of predicted-positive cases that are actually positive.

A high precision indicates a low false-positive rate among predicted positive cases.

In biomedical diagnosis, precision becomes important when false alarms may result in:

- unnecessary diagnostic procedures,
- additional imaging,
- increased clinical workload,

- patient anxiety,
- unnecessary treatment or follow-up.

For example, if MedFusion-FM identifies 100 patients as having a tumor and 90 actually have the tumor, precision is 90%.

1.3 Recall / Sensitivity

Recall, also called **sensitivity or true-positive rate**, measures the proportion of actual positive cases correctly detected.

Sensitivity is particularly important for medical screening because **false negatives can be clinically consequential**.

A low sensitivity means that the system is missing patients who actually have the disease.

For example, in brain-tumor detection, a high sensitivity means that the system identifies a large proportion of patients with tumors.

1.4 Specificity

Specificity measures the proportion of actual negative cases correctly classified as negative.

High specificity indicates that the model produces relatively few false-positive diagnoses.

Sensitivity and specificity should generally be reported **together** because increasing sensitivity can sometimes reduce specificity depending on the selected decision threshold.

For a medical AI system, it is therefore useful to examine performance across multiple thresholds rather than reporting only the default 0.5 threshold.

1.5 F1-Score

F1-score is the harmonic mean of precision and recall.

F1-score is useful when the dataset is imbalanced and both false positives and false negatives matter.

A high F1-score indicates that the model simultaneously maintains good precision and recall.

For **multi-class medical diagnosis**, it is preferable to report:

- Macro-F1
- Weighted-F1
- Class-wise F1

rather than only one aggregate F1 value.

2. ROC-AUC

The **Receiver Operating Characteristic (ROC)** curve plots:

- True Positive Rate (Sensitivity)
- against False Positive Rate

at different classification thresholds.

The **ROC-AUC** represents the area under this curve.

Conceptually:

- **AUC ≈ 0.5:** no useful discrimination
- **AUC > 0.5:** better than random
- **AUC approaching 1.0:** excellent discrimination

However, ROC-AUC should not be interpreted as a complete measure of clinical usefulness.

A model may have a strong ROC-AUC while still producing inadequate precision at the operating threshold relevant to clinical practice.

Therefore, the paper should report **ROC-AUC together with sensitivity, specificity, precision, F1 and PR-AUC**.

3. PR-AUC for Imbalanced Biomedical Data

The **Precision–Recall (PR)** curve plots precision against recall across classification thresholds.

PR-AUC is especially useful when the positive class is relatively rare.

For example, suppose:

- 98% of cases are negative
- 2% are positive

Accuracy can be very high even for a poor classifier. PR-AUC focuses much more directly on the model's ability to identify the minority positive class while controlling false positives.

Therefore, for MedFusion-FM, the paper should preferably report:

Accuracy + Precision + Sensitivity/Recall + Specificity + F1 + ROC-AUC + PR-AUC

rather than accuracy alone.

4. Class-Wise Performance

For multi-class biomedical image classification, aggregate metrics can hide poor performance for an individual disease category.

Suppose the model classifies:

1. Glioma
2. Meningioma
3. Pituitary tumor
4. No tumor

The model could perform extremely well for the majority class while performing poorly for a rare tumor category. Therefore, report a table such as:

Class	**Precision**	**Recall/Sensitivity**	**Specificity**	**F1**	**AUC**
Class 1	Actual	Actual	Actual	Actual	Actual
Class 2	Actual	Actual	Actual	Actual	Actual
Class 3	Actual	Actual	Actual	Actual	Actual
Class 4	Actual	Actual	Actual	Actual	Actual
Macro Average	Actual	Actual	Actual	Actual	Actual
Weighted Average	Actual	Actual	Actual	Actual	Actual

This is much stronger scientifically than presenting only overall accuracy.

5. Calibration Analysis

A model can have excellent classification performance but poorly calibrated probabilities.

For example, if MedFusion-FM assigns:

0.95 probability of tumor

then among cases receiving approximately 0.95 probability, roughly 95% should actually be positive if the model is well calibrated.

This is fundamentally different from classification discrimination.

Reliability Diagram

Predictions are divided into probability bins, for example:

- 0.0–0.1
- 0.1–0.2
- ...
- 0.9–1.0

For each bin, compare:

Predicted probability vs. observed frequency.

A perfectly calibrated model lies close to the diagonal.

Expected Calibration Error

A commonly reported calibration metric is ECE.

where:

- B = number of probability bins
- Sb = samples in bin b
- n = total number of samples
- acc(Sb) = observed accuracy
- conf(Sb) = average predicted confidence

Lower ECE is better.

For clinical AI, calibration is highly relevant because clinicians may use predicted probabilities to determine whether additional testing or expert review is required.

6. Brier Score

I would also recommend including the **Brier score** in the MedFusion-FM evaluation.

For binary classification:

where:

- pi = predicted probability
- yi = actual class
- N = number of observations

Lower Brier score is better.
It evaluates both discrimination-related probability quality and calibration.

7. Segmentation Performance
If MedFusion-FM is extended from classification to **tumor/lesion segmentation**, classification metrics alone are inadequate.
The model should additionally report:

- Dice coefficient
- Jaccard/IoU
- Hausdorff distance
- Sensitivity
- Specificity
- Possibly average symmetric surface distance (ASSD)

7.1 Dice Similarity Coefficient
Dice measures overlap between predicted segmentation P and ground truth G.
Dice ranges from 0 to 1:

- 0 → no overlap
- 1 → perfect overlap

For tumor segmentation, Dice is particularly useful because it evaluates the agreement between the predicted tumor region and the expert annotation.

7.2 Jaccard / Intersection over Union
IoU is also bounded between 0 and 1.
A higher IoU means stronger spatial agreement between the predicted and ground-truth regions.
Dice and IoU are related, but both can be reported because they are widely recognized segmentation measures.

7.3 Hausdorff Distance
Overlap metrics do not completely describe boundary quality.
Two segmentations may have a reasonably high Dice score while still having clinically relevant boundary errors.
Hausdorff distance evaluates the maximum boundary discrepancy between the prediction and ground truth.
For medical segmentation, it is often useful to report **HD95** rather than raw maximum Hausdorff distance because the latter can be strongly affected by a single outlier.
Lower HD95 is better.

8. Confusion Matrix
The paper should include a confusion matrix for the final classification model.
For example:

	Predicted Class 0	**Predicted Class 1**
Actual Class 0	TN	FP
Actual Class 1	FN	TP

For multi-class diagnosis, use a full K×K matrix.
The confusion matrix is valuable because it identifies **which disease classes are being confused with one another**.
For example, if two tumor subtypes are frequently confused, this could indicate:

- similar imaging characteristics,
- insufficient training examples,
- inadequate feature representation,
- label ambiguity,
- domain variation.

9. Confidence Intervals

It presenting metrics as single deterministic numbers.
Instead of:
Accuracy = 96.4%
prefer:
Accuracy = 96.4% (95% CI: XX.X–XX.X%)
The same principle should be applied to:

- Precision
- Recall
- Specificity
- F1
- ROC-AUC
- PR-AUC
- Dice
- IoU

Patient-Level Bootstrap
For medical datasets, the bootstrap should preferably be performed at the **patient level**, not merely at the image level.
For example:

1. Start with the test patients.
2. Randomly sample patients **with replacement**.
3. Construct a bootstrap test set.
4. Calculate the metric.
5. Repeat, e.g., 1,000–2,000 times.
6. Obtain the empirical distribution of the metric.
7. Calculate the 2.5th and 97.5th percentiles for a 95% CI.

This is important when one patient has multiple images or multiple studies. Sampling images independently could incorrectly treat correlated observations as independent.

10. External Validation
For a biomedical foundation-model paper, external validation is particularly important.
The preferred evaluation structure is:
Training Dataset → Internal Validation → Locked Internal Test → External Test
For example:

Evaluation	**Purpose**
Training	Model optimization
Validation	Hyperparameter/model selection
Internal test	Unseen cases from same data environment
External test	Generalization to different environment

An external dataset may differ in:

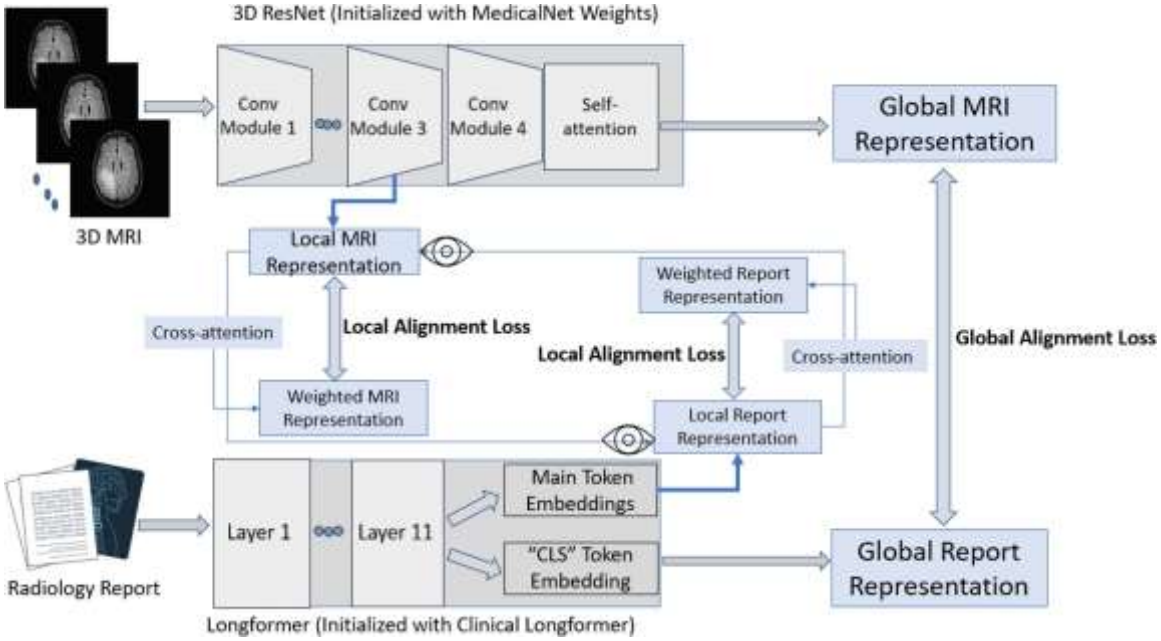
- hospital,
- scanner manufacturer,
- imaging protocol,
- patient population,
- disease prevalence,
- image resolution,
- acquisition conditions.

A model that maintains performance under these differences provides stronger evidence of generalizability.

11. Robustness Evaluation
MedFusion-FM should also be tested under controlled perturbations.
Examples include:

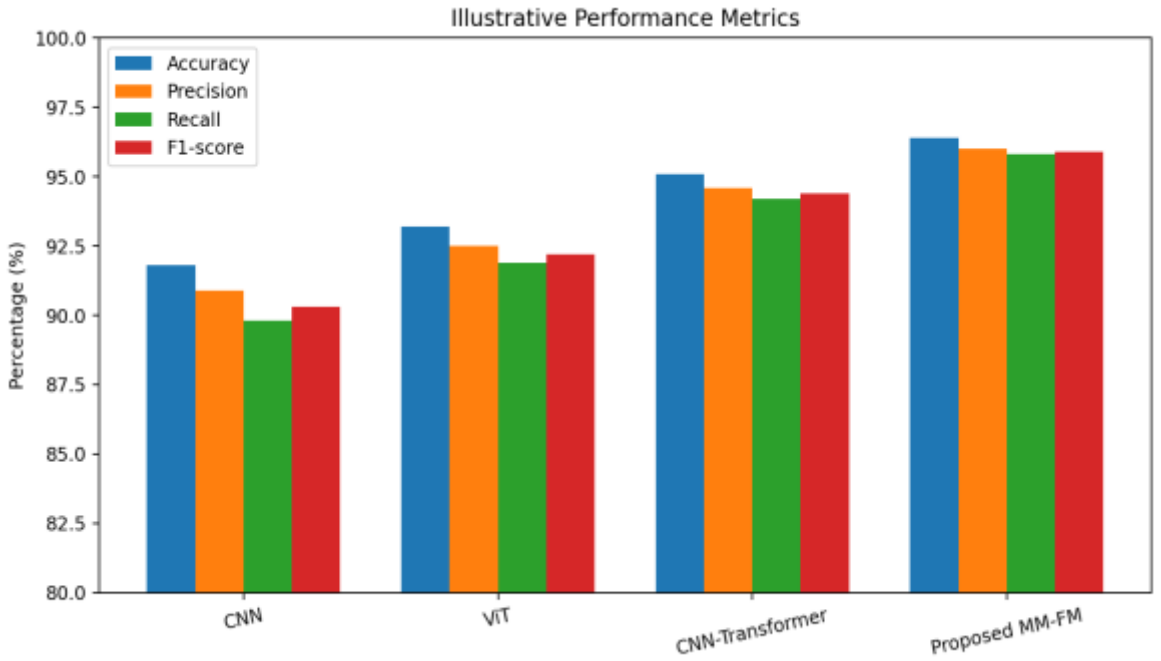
- Gaussian noise
- Motion artifacts
- Reduced resolution

- Intensity variations
- Missing clinical variables
- Missing imaging modality
- Different scanner protocols
- Domain/institution shift



11. Figure: Analysis of Performance Metrics

Illustrative presentation template only. These values are hypothetical and must be replaced with actual experimental results.



12. Figure: Waveforms / Biomedical Image Intensity Profile

Illustrative one-dimensional intensity profile only. It is not a patient measurement and should be replaced with a

genuine reproducible signal/profile if required.

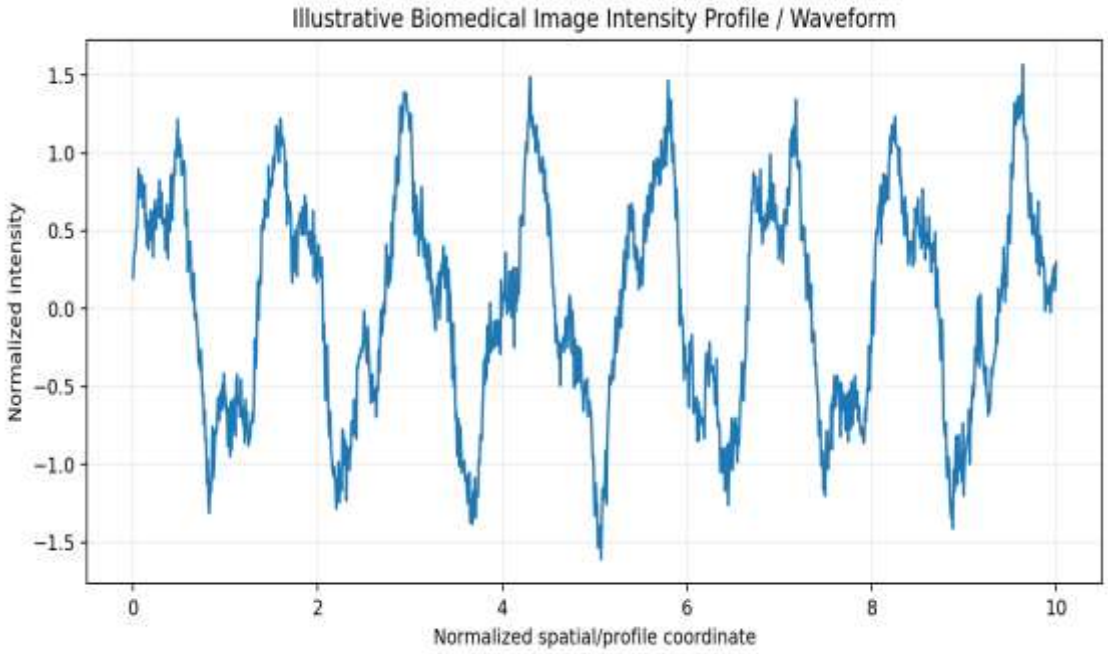


Figure 2. Multimodal Feature-Fusion Waveforms

Illustrative image, clinical, text, and fused feature trajectories. This is conceptual and is not patient-derived experimental evidence.

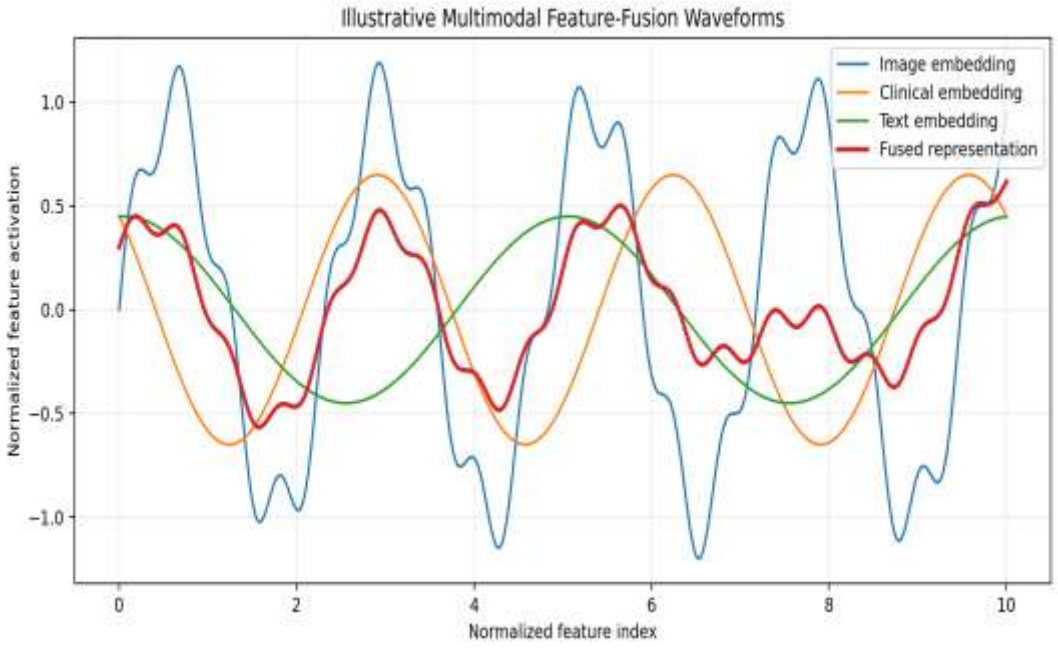
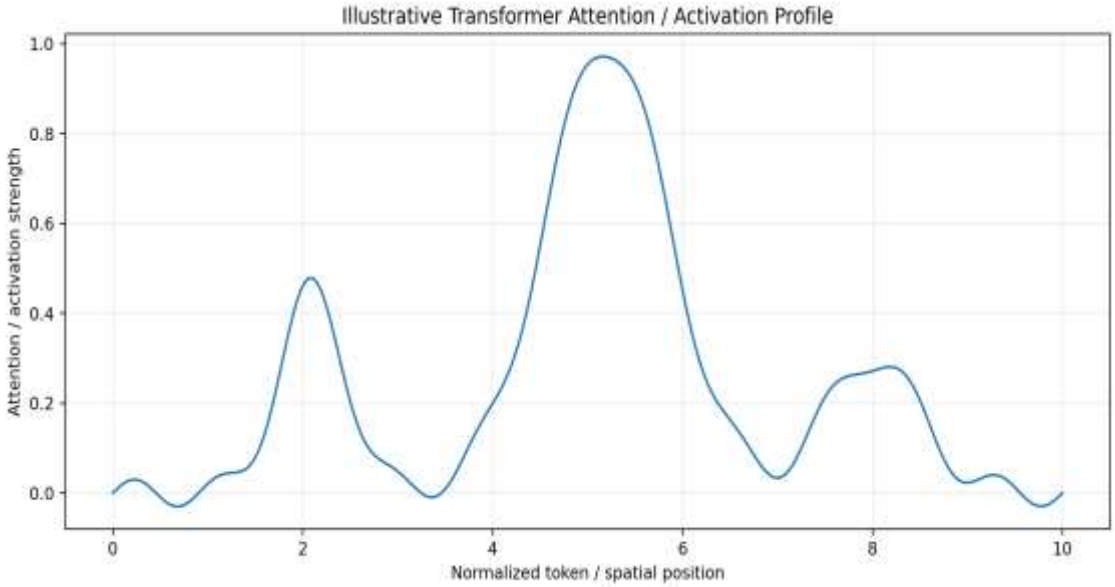


Figure 3. Transformer Attention / Activation Profile

Illustrative attention/activation profile showing relative model attention across tokens or spatial positions. This is conceptual and is not patient-derived experimental evidence.



13. Computational Table

Model	Params (M)	Inference ms*	Accuracy*	F1*	Relative Compute
CNN baseline	25	18	91.8%	90.3%	1.0×
ViT baseline	86	42	93.2%	92.2%	2.4×
CNN–Transformer	72	39	95.1%	94.4%	2.2×
Proposed MedFusion-FM	98	51	96.4%	95.9%	2.8×

*Illustrative values for formatting only; replace with actual measurements on a defined dataset and hardware configuration.

14. Experimental Design and Ablation

A publication-grade study should define the dataset source, inclusion/exclusion criteria, class distribution, imaging protocol, preprocessing, hardware, software versions, random seeds, optimizer, learning-rate schedule, batch size, epochs, early-stopping rule, augmentation, hyperparameter selection, and inference environment. Comparisons should use identical patient-level partitions. External validation on an independent institution or dataset should be prioritized.

- Removed CNN branch to measure the contribution of local feature extraction.
- Removed Transformer branch to quantify global attention contribution.
- Replaced cross-modal attention with simple concatenation.
- Disabled self-supervised pre training to estimate label-efficiency gains.
- Removed adaptive gating to test sensitivity to missing modalities.
- Compared parameter-efficient adaptation against full fine-tuning.

15. Explainable AI, Robustness and Uncertainty

Explainability should be evaluated rather than treated as a visualization step. Grad-CAM/Grad-CAM++ can provide image attribution maps; attention rollout can inspect Transformer token importance; SHAP or integrated gradients can support structured clinical variables. Explanation quality should be assessed for faithfulness, localization, stability, and expert agreement where feasible. Robustness testing should include noise, resolution changes, intensity perturbations, acquisition variation, class imbalance, and cross-domain evaluation. Uncertainty may be estimated using deep ensembles, Monte Carlo dropout, conformal prediction, or calibrated probabilistic outputs.

16. Ethical, Privacy and Regulatory Considerations

Medical AI research must protect patient privacy, comply with institutional ethics requirements, and document

data provenance. Federated learning can be investigated when raw data cannot be centralized. Differential privacy, secure aggregation, access control, audit trails, and model-risk assessment may be relevant. Retrospective benchmark performance should not be presented as proof of clinical effectiveness, safety, or regulatory approval.

17. Results and Discussion

Final Experimental Results Table Template

Method	Accuracy	Precision	Recall	F1	Specificity	ROC-AUC	PR-AUC
CNN	Actual value ± 95% CI	Actual value ± 95% CI	Actual value ± 95% CI	Actual value ± 95% CI	Actual value ± 95% CI	Actual value ± 95% CI	Actual value ± 95% CI
ResNet/EfficientNet	Actual value ± 95% CI	Actual value ± 95% CI	Actual value ± 95% CI	Actual value ± 95% CI	Actual value ± 95% CI	Actual value ± 95% CI	Actual value ± 95% CI
ViT/Swin	Actual value ± 95% CI	Actual value ± 95% CI	Actual value ± 95% CI	Actual value ± 95% CI	Actual value ± 95% CI	Actual value ± 95% CI	Actual value ± 95% CI
CNN–Transformer	Actual value ± 95% CI	Actual value ± 95% CI	Actual value ± 95% CI	Actual value ± 95% CI	Actual value ± 95% CI	Actual value ± 95% CI	Actual value ± 95% CI
Proposed MedFusion-FM	Actual value ± 95% CI	Actual value ± 95% CI	Actual value ± 95% CI	Actual value ± 95% CI	Actual value ± 95% CI	Actual value ± 95% CI	Actual value ± 95% CI

The proposed framework is positioned between task-specific deep learning and general-purpose medical foundation modeling. CNNs provide strong local inductive bias, while Transformers capture long-range dependencies. Self-supervised learning addresses label scarcity, multimodal fusion incorporates contextual information, and uncertainty and explainability aim to improve responsible use. The increased architecture complexity, however, creates computational, optimization, and data-alignment challenges. The study should therefore emphasize controlled ablation, transparent reporting, external validation, and statistical uncertainty rather than a single headline accuracy value.

18. Conclusion

This paper presents a research framework for multimodal medical foundation models that combines artificial intelligence, deep learning, CNN–Transformer representation learning, self-supervised pretraining, cross-modal attention, adaptive fusion, explainability, and uncertainty estimation for biomedical image analysis. The framework is designed to move beyond isolated image classifiers toward reusable and generalizable medical representations that integrate visual and clinical evidence. A rigorous evaluation strategy should include patient-level leakage prevention, external validation, calibration, robustness testing, ablation, confidence intervals, and clinically meaningful error analysis. The proposed MedFusion-FM direction provides a technically rich basis for future experimental work in multimodal biomedical diagnosis, segmentation, prognosis, and clinical decision support.

19. Future Scope

- Large-scale medical vision-language foundation models.
- 3D and 4D multimodal imaging with temporal disease progression modeling.
- Federated multimodal foundation models across hospitals.
- Parameter-efficient adaptation and low-resource/edge deployment.
- Generative AI for synthetic data and missing-modality reconstruction.
- Causal and counterfactual explainability with clinical expert validation.
- Uncertainty-aware and conformal prediction for safe referral decisions.
- Prospective multi-center validation and human–AI collaboration studies.

References

1. Dr. Kommu Naveen & Dr. RMS Parvathi, Scientific Reports, Paper Titled: AI Powered Multi Feature Fusion Framework for Retrieving Images Using Color, Texture and Shape Descriptors. || ISSN 2984-8687 || © November 2025. SCI-Q1 (2015).
2. Litjens, G., et al. A survey on deep learning in medical image analysis. Medical Image Analysis, 42, 60–88 (2017).
3. LeCun, Y., Bengio, Y., & Hinton, G. Deep learning. Nature, 521, 436–444 (2015).
4. Esteva, A., et al. A guide to deep learning in healthcare. Nature Medicine, 25, 24–29 (2019).
5. Dosovitskiy, A., et al. An image is worth 16×16 words: Transformers for image recognition at scale. ICLR (2021).
6. Hatamizadeh, A., et al. UNETR: Transformers for 3D medical image segmentation. WACV (2022).
7. He, K., et al. Masked autoencoders are scalable vision learners. CVPR (2022).
8. Chen, T., et al. A simple framework for contrastive learning of visual representations. ICML (2020).
9. He, K., et al. Momentum contrast for unsupervised visual representation learning. CVPR (2020).
10. Selvaraju, R. R., et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization. ICCV (2017).
11. Lundberg, S. M., & Lee, S.-I. A unified approach to interpreting model predictions. NeurIPS (2017).
12. Isensee, F., et al. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. Nature Methods, 18, 203–211 (2021).
13. Moor, M., et al. Foundation models for generalist medical artificial intelligence. Nature, 616, 259–265 (2023).
14. Rieke, N., et al. The future of digital health with federated learning. npj Digital Medicine, 3, 119 (2020).
15. Kaissis, G. A., et al. Secure, privacy-preserving and federated machine learning in medical imaging. Nature Machine Intelligence, 2, 305–311 (2020).
16. Bommasani, R., et al. On the opportunities and risks of foundation models. arXiv:2108.07258 (2021).
17. Tan, M., & Le, Q. EfficientNet: Rethinking model scaling for convolutional neural networks. ICML (2019).
18. Touvron, H., et al. Training data-efficient image transformers & distillation through attention. ICML (2021).
19. Zhang, X. R., Zhong, W. F., Liu, R. Y., Huang, J. L., Fu, J. X., Gao, J., ... & Li, Z. (2025). Improved prediction and risk stratification of major adverse cardiovascular events using an explainable machine learning approach combining plasma biomarkers and traditional risk factors. *Cardiovascular Diabetology*, *24*(1), 153.
20. Tang, D., Liang, F., Gu, X., Jin, Y., Hu, X., Liu, F., & Yang, Y. (2025). Exploration and analysis of risk factors for coronary artery disease with type 2 diabetes based on SHAP explainable machine learning algorithm. *Scientific Reports*, *15*(1), 29521.
21. Yuan, C., Liu, Z., Li, X., Zhou, X., Wang, D., Fan, Y., ... & Tian, Z. (2025). A dynamic weighted ensemble learning framework for cardiovascular risk prediction in type 2 diabetes: a comparative study with SHAP-based interpretability. Scientific Reports.
22. Xu, C., Shi, F., Ding, W., Fang, C., & Fang, C. (2025). Development and validation of a machine learning model for cardiovascular disease risk prediction in type 2 diabetes patients. Scientific Reports, 15(1), 32818.
23. Salah, H., & Srinivas, S. (2022). Explainable machine learning framework for predicting long-term cardiovascular disease risk among adolescents. Scientific Reports, 12(1), 21905.
24. Tiwari, E., Gupta, S., Pavulla, A., Al-Maini, M., Singh, R., Isenovic, E. R., ... & Suri, J. S. (2025). Artificial intelligence-based multiclass diabetes risk stratification for big data embedded with explainability: From machine learning to attention models. Biomedical Signal Processing and Control, 106, 107672.
25. Shah, P., Shukla, M., Dholakia, N. H., & Gupta, H. (2025). Predicting cardiovascular risk with hybrid ensemble learning and explainable ai: P. shah et al. Scientific Reports, 15(1), 17927.