Journal of Information Systems Engineering and Management
2025, 10(12s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Multimodal Federated Transformer: Advancing Medical Image Captioning with MASNet-ViT Fusion and Enhanced Spiking Neural Networks

Mrs. AVS Ratna Kumari[1], Dr. Lalitha Kumari Pappala[2,*]

[1] *Research Scholar, School of Computer Science and Engineering, VIT-AP University, Amaravati, 522237, India.*
Email: ratna.23phd7109@vitap.ac.in
[2] *School of Computer Science and Engineering, VIT-AP University, Amaravati, Guntur, 522237, India.*
Email: lalitha.p@vitap.ac.in
**Corresponding author: Dr Lalitha Kumari Pappala (Email:lalitha.p@vitap.ac.in)*

ARTICLE INFO

Received: 18 Nov 2024

Revised: 30 Dec 2024

Accepted: 18 Jan 2025

ABSTRACT

Determining the density, coefficient of thermal expansion of BaTiO3 across range of 400 K to 1075 K is principal goal of the investigation. Coefficient of temperature dependence of density and coefficient of volume thermal expansion of the compound is assessed. Additionally, the research has been broadened to evaluate the linear attenuation coefficient at various γ-energies over a temperature range using value of mass attenuation coefficient.

Keywords: Medical Image Captioning, MASNet-ViT Fusion, Enhanced Spiking Neural Networks, Federated Learning, Diagnostic Automations.

Abbreviation	Full Form
MFT	Multimodal FederatedTransformer
MASNet	Multiscale Attention Network
ViT	Vision Transformer
ESNN	Enhanced Spiking Neural Network
FL	Federated Learning
MIMIC-CXR	Medical Information Mart for Intensive Care Chest X-rays
IU X-ray	Indiana University Chest X-ray Dataset
PEIR	Pathology Education Instructional Resource
BLEU	Bilingual Evaluation Understudy
CIDEr	Consensus-based Image Description Evaluation
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
HIPAA	Health Insurance Portability and Accountability Act
XAI	Explainable Artificial Intelligence
EHR	Electronic Health Records
PET	Positron Emission Tomography
CT	Computed Tomography
ms	Milliseconds
hrs	Hours

I. INTRODUCTION

Medical image analysis plays an indispensable role in modern diagnostics, providing critical insights into various pathologies. The major difficulty remains in transforming complex imaging data into meaningful clinical interpretations, and this is most challenging in a resource-constrained setting. Conventional image captioning models [1, 2, 3] were highly effective for natural images, but they failed to satisfy the peculiar requirements of medical imaging owing to the intricacy of anatomy, variety in imaging modalities, and the subtle features of pathology. In addition, centralization of data aggregation in traditional deep learning approaches often leads to privacy issues and affects generalization across different healthcare institutions in process. Some of the above challenges have recently been addressed through advancements in deep learning, especially with vision transformers and multiscale attention networks. MASNet is particularly effective at discovering fine local features, while ViTs are successfully extracting global dependencies and contextual patterns. So far, most of the approaches that apply such techniques are often inadequate to provide an all-round understanding of the medical images & samples. At the same time, caption generation models generally rely on dense representations, which may not be suitable for sparse and temporal medical data samples. Overcoming these challenges [4, 5, 6], this paper proposes the Multimodal Federated Transformer (MFT) framework, integrating MASNet-ViT fusion, Enhanced Spiking Neural Networks (ESNNs), and Federated Learning (FL). The weighted aggregation mechanism is used to fuse MASNet and ViT features, so that the framework is balanced between local and global insights, which enhances its capability to detect subtle abnormalities and holistic patterns. The ESNN module uses biologically inspired spiking neurons for dynamic temporal processing, offering efficient and precise caption generations. In addition, FL allows decentralized model training across several healthcare institutions with data privacy and compliance with regulations such as HIPAA while improving model generalization operations. This paper demonstrates the effectiveness of the MFT framework through comprehensive experiments on diverse datasets, showing superior captioning metrics and diagnostic impact sets. It fills the gap between advanced image understanding and practical clinical applications, providing a new benchmark in the process of automated medical image captioning process

Motivation and Contribution

The increasing dependence on medical imaging for diagnostic workflows underlines the huge need for more automated tools that can generate contextually relevant and accurate captions. Most of the current methods for medical image captioning fail to meet the precision and adaptability demands of clinical utilization due to their failure to simultaneously capture localized abnormalities and broader contextual patterns in complex images & samples. In addition, the traditional centralized training methods raise issues of ethical and regulatory concerns on patient data privacy, especially considering the strict regulations such as HIPAA and GDPR sets. The above limitations clearly indicate a gap in the ability to use state-of-the-art machine learning methodologies for secure, scalable, and effective medical image captioning process.

The proposed work introduces a novel solution, namely the Multimodal Federated Transformer framework, to tackle these challenges. There are three main contributions provided by this framework. Firstly, it introduces a fusion that includes MASNet and Vision Transformer. This, with a weighted aggregation mechanism, would provide an adequate balance of extracting local details versus understanding contextual awareness globally. It utilizes ESNN for generating captions based on dynamic temporal spikes in order to ensure efficient biologically inspired sparse data representation. It uses a Federated Learning (FL) framework that allows decentralized model training across institutions while preserving the privacy of patient sets. Evaluation on benchmark datasets shows that MFT outperforms the state-of-the-art methods significantly in terms of quality of captions, diagnostic accuracy, and scalability levels. This work adds a robust and adaptable, clinically impactful solution for automated medical image interpretation sets by addressing critical gaps in the following areas: feature extraction, caption generation, and data privacy.

II. LITERATURE REVIEW

Reviewing the image captioning methodologies based on recent paper analysis enlightens a broad range of developments, challenges, and innovations in this domain. This corpus covers medical, artistic, and construction-related imagery and includes diverse techniques such as transformers, generative adversarial networks, recurrent neural networks, and multimodal approaches. These studies taken together emphasize image captioning with tremendous potential on real-world applicability in automatically enriching documentations, diagnoses, and even contextual understanding operation. The domains for domain-specific image captioning discussed in [1] and in Sharma's extensive review demonstrate challenges that face general models, which have to specifically fit domain specifics, thus raising a need to be tailored from a general standpoint but adapted towards handling various domain-specific niceties. Selivanov et al. [2] pushed the applicability of pretraining transformers in a generative sense for medical image captioning. They attained excellent fluency and contextual appropriateness for the captions that emerged as a consequence of this activity. The work by Selivanov et al. is also quite close to that of Sharma and Padha who employed Neuraltalk+ with visual aid to get improved semantic matches [3]. Fine-grained captioning was elevated to the next dimension when GANs were proposed by Yang et al. [4]. For instance, the methodology adopted by the authors ensures that captioning captures nuances in emotion and context, which is important for applications such as art and media collections. In similar lines, Ren et al. [5] developed cross-attention transformers to improve the alignment between visual and textual modalities. A survey from Salgotra et al. [6] situates this progress within a broader context by pointing to the emerging trends of multimodal fusion and attention-based mechanisms.

TABLE I. METHODOLOGICAL EMPIRICAL REVIEW ANALYSIS

Reference	Method	Main Objectives	Findings	Limitations
[1]	Domain-specific image captioning review	Comprehensive review of image captioning in specific domains	Identified domain-specific challenges and highlighted key techniques	Limited coverage of emerging techniques such as transformers
[2]	Generative pretrained transformers (GPT) for medical image captioning	Improve medical image captioning using GPT models	Achieved higher fluency and contextual relevance in captions	High computational cost and limited adaptability to small datasets
[3]	Neuraltalk+	Neural captioning with visual assistance	Enhanced semantic alignment and accuracy in captions	Struggles with highly complex image contexts
[4]	Fine-grained image emotion captioning using GANs	Capture emotional context in captions	Improved contextual sensitivity for artistic and media domains	Requires extensive labeled data for GAN training
[5]	Cross-attention-based image captioning transformer	Improve visual-textual modality alignment	Achieved state-of-the-art accuracy for general image captioning tasks	Computationally expensive and domain-agnostic
[6]	Survey on automatic image captioning	Overview of trends and future directions in image captioning	Identified key advancements in multimodal fusion and attention mechanisms	Lack of experimental insights into specific techniques
[7]	Self-Enhanced Attention (SEA)	Improved attention mechanisms for image captioning	Better handling of long-range dependencies in captions	May overfit on datasets with limited variability
[8]	Concept-based LSTM and multi-encoder transformer	Introduce novel architectures for image captioning	Enhanced semantic understanding for complex datasets	Performance dependent on hyperparameter tuning
[9]	Unified multitask learning model	Combine classification, detection, and captioning	Efficient multitask representation learning	Model complexity increases training time
[10]	Dynamic text prompt multimodal features	Captioning for plant disease images	Achieved higher accuracy with joint features	Limited generalizability beyond agricultural domains
[11]	Augmentation and ranking mechanism	Improve automatic image captioning	Enhanced robustness across datasets	Limited scalability to highly complex

		systems		datasets
[12]	Improved Arabic image captioning	Use pre-trained word embeddings for Arabic captions	Improved linguistic fluency and contextual relevance	Focused solely on Arabic, limiting cross-linguistic insights
[13]	Descriptive captioning for histopathological patches	Enhance captioning in pathology	Generated precise and informative captions for histopathology	Dataset-dependent performance
[14]	Multimodal feature fusion with mask RNN and LSTM	Improve multimodal fusion in captioning	State-of-the-art performance on multimodal datasets	High computational requirements
[15]	Captioning for cultural artifacts	Domain-specific captioning for ceramics	Achieved detailed and contextually accurate captions	Limited to cultural artifacts
[16]	Prior-knowledge transformer for ECG captioning	Incorporate domain knowledge into captions	Improved diagnostic relevance in medical reports	Specific to ECG data, limiting general applicability
[17]	IQAGPT	Use GPT for image quality assessment in CT	Improved automated quality assessment	Requires large-scale datasets for effective performance
[18]	Comprehensive review of image caption generation	Overview of advancements in image captioning	Highlighted the evolution of neural architectures	Lack of focus on emerging multimodal datasets
[19]	Multimodal transformer for medical image analysis	Automated report generation for medical images	Enhanced integration of visual and textual data	High computational complexity
[20]	Medtransnet	Gating transformer for medical classification	Improved accuracy in medical diagnostics	High dependency on labeled data
[21]	Clustering swap prediction	Enhance image-text pretraining	Improved cross-modal alignment	Limited scalability to diverse datasets
[22]	Survey on datasets and methods	Comprehensive review of datasets for image captioning	Identified benchmark datasets and methods	Lack of experimental validation
[23]	Vision Transformers vs. CNNs	Compare transformers and CNNs for medical imaging	Highlighted strengths of transformers in feature extraction	Lack of analysis on hybrid models
[24]	Generative foundation model	Self-improving models for synthetic image generation	Enhanced data augmentation for medical applications	High resource requirements
[25]	Survey on medical imaging report generation	Review of report generation techniques	Explored deep learning approaches for clinical reporting	Limited coverage of emerging large-scale models
[26]	Dense deep transformer (DDTraMIS)	Transformer-based segmentation for	Improved segmentation accuracy	High model complexity

		medical images		
[27]	Multi-expert fusion network (MeFD-Net)	Diagnostic network for radiology image reports	Enhanced diagnostic accuracy through fusion	Performance bottlenecks on large-scale datasets
[28]	Hybrid attention with Laplacian query fusion	Improve medical image segmentation	Better handling of complex segmentation tasks	Requires fine-tuning for specific datasets
[29]	Self-supervised learning review	Guidelines for medical image classification	Provided implementation insights for self-supervised techniques	Lack of focus on specific applications like captioning
[30]	Encoder-decoder for automated captioning	Improve efficiency in captioning	Achieved competitive performance across datasets	Limited handling of contextual nuances
[31]	Sentiment-based cues for image classification	Use linguistic cues to aid classification	Enhanced contextual understanding in captions	Focused on sentiment analysis, limiting general applicability
[32]	Central attention with multi-graphs	Improved annotation through graph-based attention	Better feature representation for annotations	Requires complex graph construction
[33]	Multimodal fusion for visual QA	Enhance visual question answering	Improved multimodal representation	Limited generalizability to other tasks
[34]	IMAD: Image-Augmented Dialogue	Multi-modal dialogue systems with image support	Enhanced interaction in multimodal settings	Limited dataset availability
[35]	Transformer-based report generator	Automatic medical report generation	Achieved high fluency and diagnostic relevance	High dependency on large datasets
[36]	BangleFIR	Fashion image retrieval dataset	Enriched datasets for retrieval tasks	Focused solely on fashion domain
[37]	Cross-modal representation learning	Image-sentence retrieval	Improved retrieval accuracy with transferable features	Limited to cross-modal retrieval tasks
[38]	Visual-language foundation model	Pathology-focused multimodal integration	Improved diagnostic support for pathology	High model complexity
[39]	Reinforced interaction fusion	Radiology report generation	Enhanced integration of visual and textual data	High resource requirements
[40]	Dual-stream multi-label classification	Improve multi-label classification with feature reconstruction	Achieved high classification accuracy	High dependency on labeled data
[41]	Enriching satellite	Satellite image	Improved contextual	Specific to satellite

	annotations	annotations with keyphrases	annotations for forests	imagery
[42]	Causal reasoning for vision tasks	Apply causal reasoning to computer vision	Improved interpretability in visual tasks	Computationally intensive
[43]	Knowledge alignment for histopathology	Align concepts with whole-slide images	Improved precision in histopathology analysis	High dependency on domain-specific ontologies
[44]	Hyperspectral image classification review	Techniques and challenges in hyperspectral imaging	Identified gaps and future directions	Lack of focus on captioning applications
[45]	Synthetic training image generation	Generate synthetic images for railway defect detection	Improved cognition in defect detection	Specific to railway defects
[46]	Mini-InternVL	Flexible-transfer multimodal model	Achieved high performance with minimal parameters	Limited testing on diverse tasks
[47]	Stochastic gradient descent for X-ray diagnosis	Enhance optimization in medical imaging	Improved diagnostic accuracy	Requires further validation on larger datasets
[48]	Zero-shot caption inference	Pretrained models for zero-shot inference	Enabled captioning without task-specific training	Struggles with highly complex domains
[49]	Semantic scene-based captioning	Image captioning using semantic scenes	Improved semantic alignment	Limited to scene-specific contexts
[50]	Style-enhanced transformer	Captioning in construction scenes	Improved fluency and context awareness	Focused solely on construction domains

Innovations in medical imaging, such as those by Sun et al. [7], and Osman et al. [8], use better attention mechanisms and multi-encoder architectures to improve the accuracy and descriptiveness of captions. Generalizing on this, Bayisa et al. [9] present a unified framework that introduces multitask learning to further evidence its ability in handling tasks such as classification, object detection, and captioning all at once in process. This is comparable with Elbedwehy et al. [12, 13], where better feature representations enhanced Arabic and histopathological image captioning pre-trained word embedding and descriptive models for captioning. Their work significantly identifies feature fusion as a key role in the captioning process, especially a semantically relevant caption that could be produced by a domain-specific scenario. Further elevating the abilities of multimodal fusion techniques, Thangavel et al. [14] applied a mask RNN and LSTMs for image captioning that found the state-of-the-art performance. Meanwhile, Zheng et al. [15] concentrated their efforts into describing the cultural artefact's such as ceramics using contextual embeddings in crafting domain-specific captions. This has been taken a notch higher by prior knowledge transformers developed by Tran et al. [16], which is an incorporation of domain knowledge in caption generation for images based on ECG, a critical innovation in the health care sets. Works like IQAGPT by Chen et al. [17] merge vision-language models with large transformers, as in the case of computed tomography evaluation using ChatGPT. It is a follow-up of the more extensive survey on transformer-based architectures for image captioning surveyed by Arshi and Dadure [18]. Their survey highlights from RNNs to transformers, depicting the trend towards models that better understand the contextual and temporal process. The pretraining strategy presented architectures that integrated multimodal learning to automate medical report generation; therefore, a streamlined diagnostic workflow is facilitated by Raminedi et al. [19], Shaik et al. [20], and Fayou et al. [21] based on clustering swap prediction for the improvement of cross-modal alignment. Agarwal and Verma [22] have conducted an exhaustive survey on the datasets, methods, and relevance of benchmark datasets in furtherance of the area.

Very recently, the systematic reviews of the comparison such as [23] that makes a comparison of the vision transformers and the convolutional network will highlight comparative strengths for application tasks in the medical imaging setting. Wang et al. extended the self-improving generative models [24], which presents an exciting direction forward, especially in synthesizing images from scratch and applicability in augmenting data synthesis. This matches up with a survey reported by Pang et al. [25] on deep learning-based report generation which elucidates the transition from static image captioning to dynamic, clinically relevant documentations. Captions and image segmentation are put together in an example architecture presented by Joshi and Sharma's [26] dense transformer segmentation architecture, while a multi-expert fusion network is proposed by Ran et al. [27] for radiology reporting. Hybrid attention mechanisms with a combination of Laplacian query fusion and sequence matching are improved for segmentation tasks by Ekong et al. [28]. Huang et al. [29] introduces a conceptual framework on self-supervised learning, giving an overview with some guidelines for implementation that can complement captioning models. Other types of research such as Ansari and Srivastava [30] talk about encoder-decoder models while Kaur et al. [31] speaks about sentiment-based cues, increasing the linguistic as well as visual features captured by captions. Liu et al. [32] built central attention mechanisms; For example, novel datasets like BangleFIR [36] are advancing progress in retrieval and captioning tasks while enriching model training by including domain-specific contexts. Such cross-modal contributions are Yang et al. for the transferable representation framework and Lu et al. for a visual-language model on computational pathology sets [38, 37]. In work by Wang et al. [39] and Hu et al. [40], such extensions took place toward reinforcement and dual-stream mechanisms on tasks of radiology and multi-label classification. Emerging domains, as reviewed by Tejasree and Agilandeeswari in hyperspectral imaging [44], and railway defect detection, investigated by Ferdousi et al. [45], indicate versatility in captioning models across varied industries and applications. Mini-InternVL as demonstrated by Gao et al. [46] indicates strong performance in the flexible transfer multimodal models involving minimal parameters within processes. Banik [47] brings stochastic gradient descent to the world of chest X-ray diagnosis and pushes the preset models toward zero-shot inference for captioning, as does Zhang et al. [48]. Zhao et al. [49] and Song et al. [50] add further to this pool with semantic scene analysis and style-enhanced transformers, respectively in the process. The combined understanding from these papers reflects a transformative journey of image captioning, which shifts towards multimodal, domain-specific, and scalable solutions.

III. PROPOSED MODEL ANALYSIS

To overcome the issues of low efficiency & high complexity that are present in existing methods, this section discusses the design of an iterative Multimodal Federated Transformer: Advancing Medical Image Captioning with MASNet-ViT Fusion and Enhanced Spiking Neural Networks. First, by referring to figure 1, the MFT framework is devised on state-of-the-art techniques in order to mitigate the internal complications associated with the task of medical image captioning. Here, by including MASNet, combined with the use of Fusion Vision Transformer and the Enhanced Spiking Neural Network ESNN to serve for the task of captioning, plus the Federated Learning framework for developing the same, this aims for a proper, privacy-preserving and precise solution to automatically interpret medical images. All these components are optimized to tap into their respective strengths while complementing one another without being difficult to integrate. The MASNet component is designed to extract localized, fine-grained features from medical images using multiscale attention mechanisms. For an input image I, the MASNet performs multiscale feature extraction by processing patches of I at varying receptive fields. The local feature map Flocal can be represented via equation 1,

$$Flocal = MASNet(I) = \sum (ws \cdot As \cdot Is), \quad for\ s = 1\ to\ S \ldots (1)$$

Where, S is the number of scales, ws represents learnable scale weights, ‘As’ is the attention map for scale s, and Is represents the corresponding input patches.

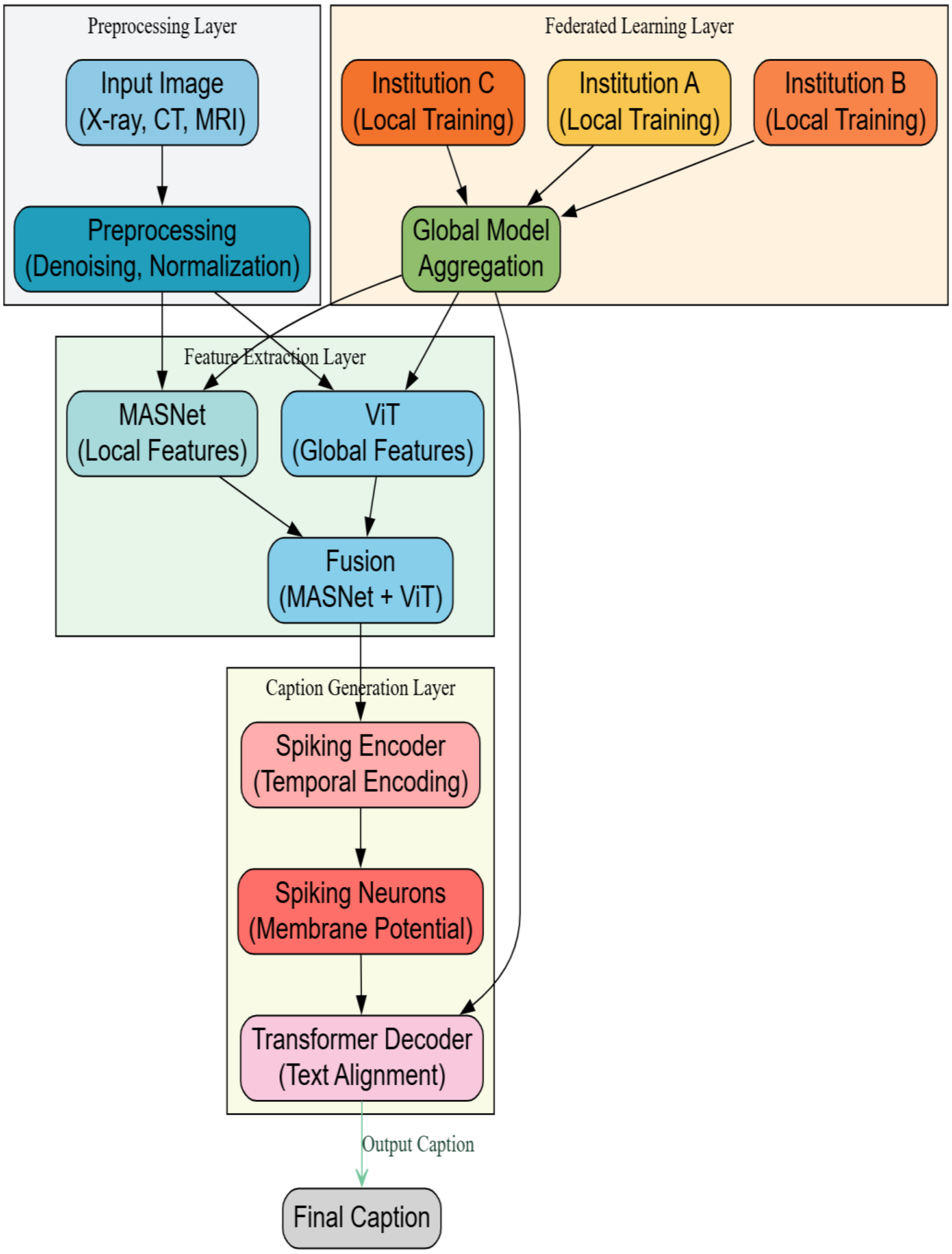
Preprocessing Layer
Input Image (X-ray, CT, MRI)
Preprocessing (Denoising, Normalization)
Federated Learning Layer
Institution C (Local Training)
Institution A (Local Training)
Institution B (Local Training)
Global Model Aggregation
Feature Extraction Layer
MASNet (Local Features)
ViT (Global Features)
Fusion (MASNet + ViT)
Caption Generation Layer
Spiking Encoder (Temporal Encoding)
Spiking Neurons (Membrane Potential)
Transformer Decoder (Text Alignment)
Output Caption
Final Caption

Fig. 1. Model Architecture of the Proposed Analysis Process

This formulation ensures that MASNet captures both the micro-level pathological detail and structural contexts. Iteratively, Next, as per figure 2, the Vision Transformer (ViT) processes the input image globally by dividing I into N non-overlapping patches. These patches are projected into an embedding space E and passed through transformer layers with self-attention process. The global feature representation Fglobal is computed via equation 2,

$$Fglobal = ViT(I) = \sum\left(an \cdot Attn(Q,K,V)\right) for\, n = 1\ to\ N \ldots(2)$$

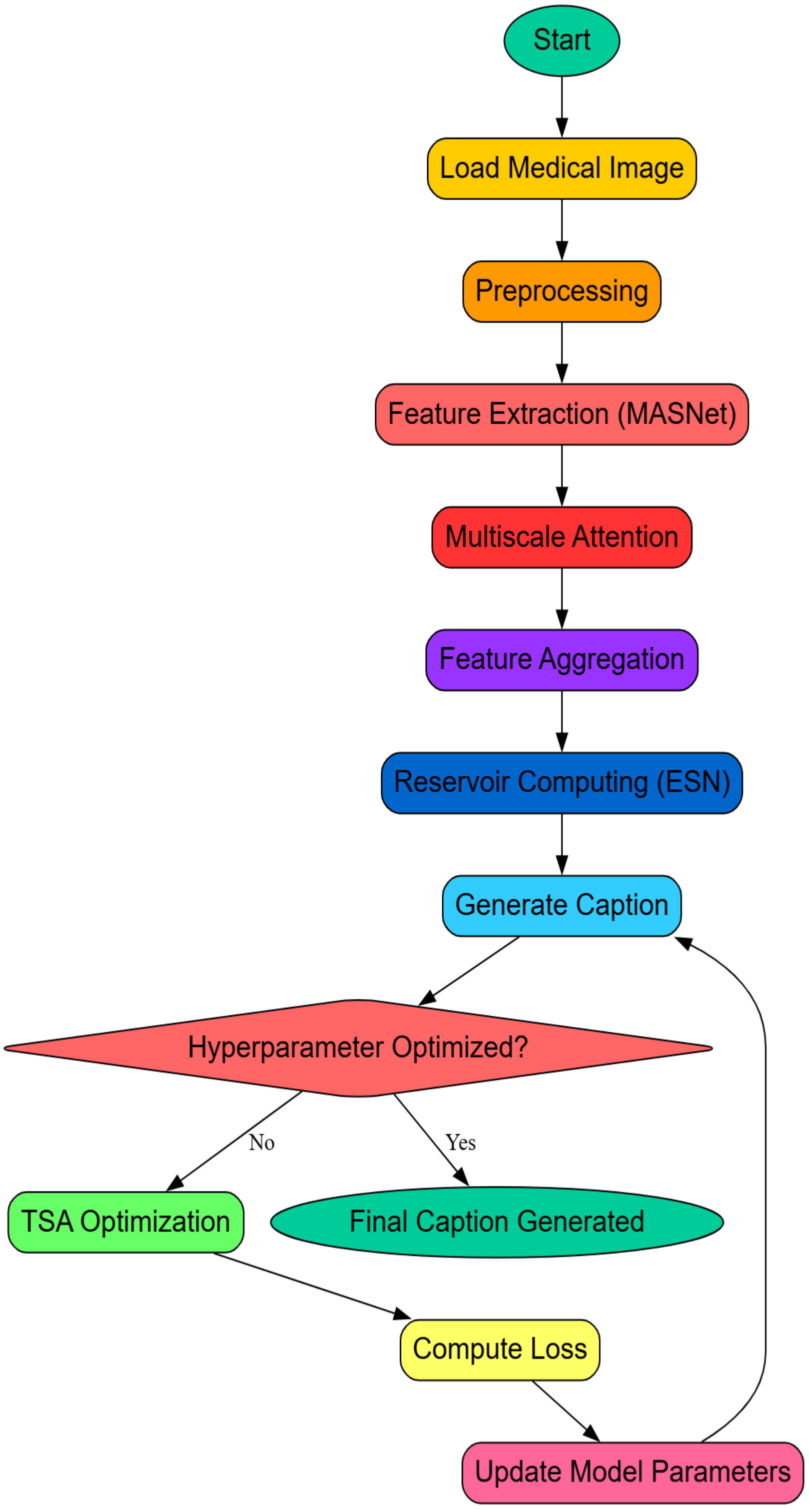

Fig. 2. Overall Flow of the Proposed Analysis Process

Where, αn are attention coefficients, and Attn(Q, K, V) represents the self-attention mechanism computed via equation 3,

$$Attn(Q,K,V) = softmax\left(\frac{QK^T}{dk}\right)V \ldots(3)$$

Where, Q, K, V are query, key and value matrices derived from the input embeddings, and dk is the dimensionality of the key vectors in this process. This ensures that ViT captures long-range dependencies and semantic contexts.

Features between MASNet and ViT are fused through weighted aggregations. Fused feature embedding Ffused is obtained via equation 4,

$$Ffused = W1 \cdot Flocal + W2 \cdot Fglobal \ldots (4)$$

Where, W1 and W2 are learned weights ensuring maximum balance between the local and the global representations. A linear projection layer refines Ffused into a feature space that makes it compact in this process. For captioning, ESNN is used during the process of the algorithm sets. The fused feature is then presented as spiking inputs by passing through an encoding layer S(t), as designed via equation 5:

$$S(t) = \frac{1}{1 + e^{-\lambda\,(Ffused\ -\theta)}} \ldots (5)$$

Where, λ is a scaling parameter, and θ is the firing threshold for this process. The encoded spikes are processed through spiking layers, governed by the membrane potential process via equation 6,

$$\frac{dU(t)}{dt} = -\frac{U(t)}{\tau} + \sum\left(wi\ Si(t)\right) \ldots (6)$$

Where the output is U(t) for membrane potential, and the synaptic weights and time constants are wi for synaptic weights and τ for membrane time constant. Here input spikes and output spikes O(t) are decoded into textual captions using a transformer decoder with an attention mechanism aligned to medical terminologies. The Federated Learning (FL) framework ensures decentralized training of MASNet-ViT and ESNN across multiple institutions in the process. Each institution 'i' locally computes gradients ∇Li for the loss function Li via equation 7,

$$Li = -\sum\left(yk\ log(\hat{y}k)\right) \ldots (7)$$

Where, yk and ŷk represent true and predicted probabilities. The global model update θt is derived using Federated Averaging (FedAvg) via equation 8,

$$\theta t = \frac{\sum\left(ni \cdot \theta(t-1)^{i}\right)}{\sum ni} for\ i = 1\ to\ N \ldots (8)$$

Where, ni represents the data size at institution 'i' in process. This aggregation ensures privacy-preserving training while improving generalization operations.The final output of the model, a contextualized medical caption, C, is represented via equation 9,

$$C = Decoder\left(O(t)\right) = softmax(Wo \cdot H + bo) \ldots (9)$$

Where, H represents the hidden state from the transformer decoder, Wo are learnable weights, and bo is the bias term for this process. This end-to-end pipeline strikes an optimal balance between interpretability, accuracy, and privacy; thus, this becomes an efficient solution for the medical image captioning process. Fig. 3 As iterated, the proposed MFT framework includes the sophisticated modules of feature extraction, caption generation, and decentralized learning to address the complex needs of the medical image captioning process. The MASNet with Vision Transformer (ViT) Fusion, Enhanced Spiking Neural Networks (ESNN), and the FL framework of its components ensure robustness, efficiency, and privacy compliance. The first part is MASNet. It extracts fine-grained local features by utilizing multiscale attention mechanisms.

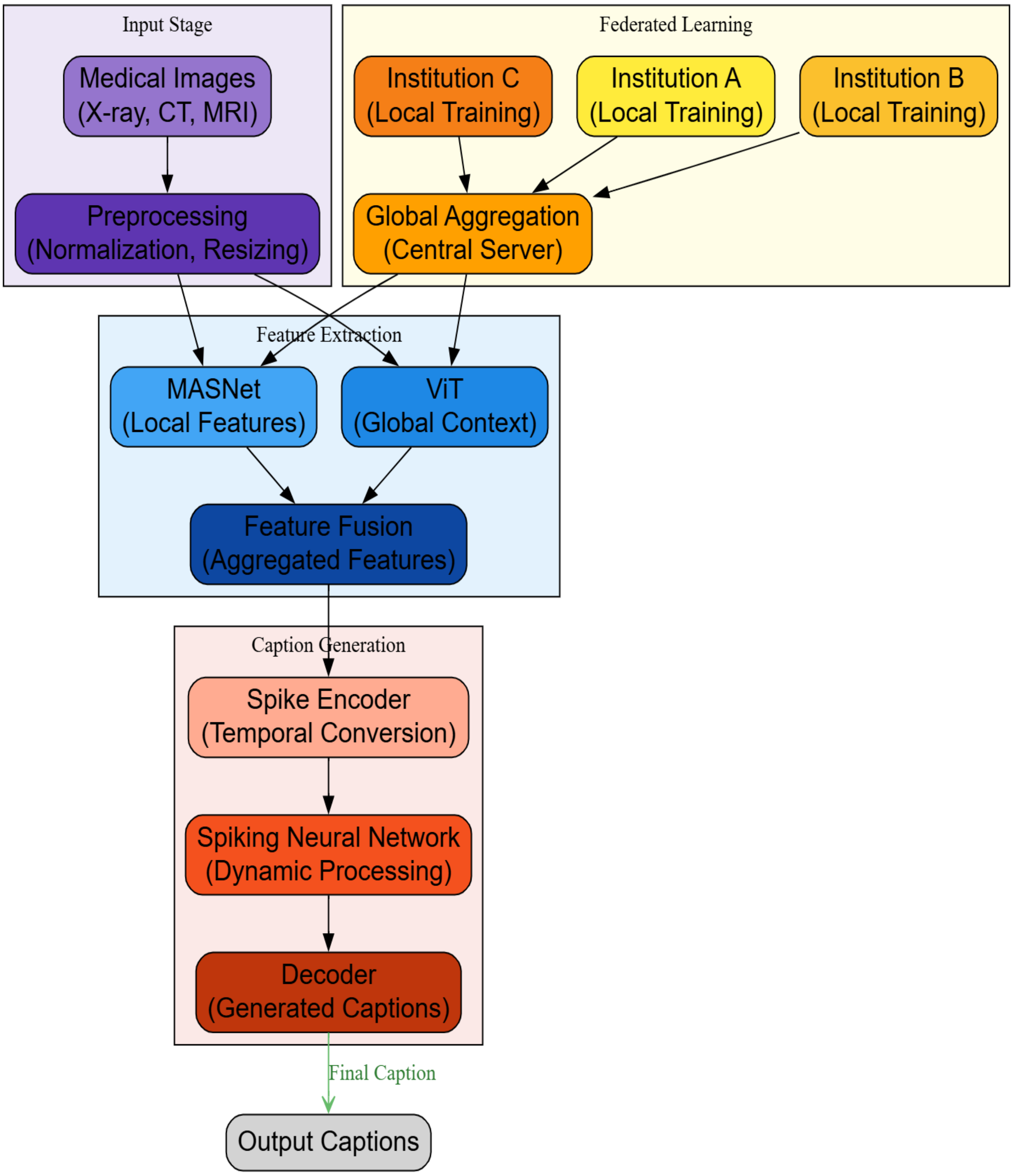
Input Stage
Medical Images (X-ray, CT, MRI)
Preprocessing (Normalization, Resizing)
Federated Learning
Institution C (Local Training)
Institution A (Local Training)
Institution B (Local Training)
Global Aggregation (Central Server)
Feature Extraction
MASNet (Local Features)
ViT (Global Context)
Feature Fusion (Aggregated Features)
Caption Generation
Spike Encoder (Temporal Conversion)
Spiking Neural Network (Dynamic Processing)
Decoder (Generated Captions)
Final Caption
Output Captions

Fig. 3. Flowchart of the Proposed Analysis Process

The attention map As is derived from learned queries, keys, and values via equation 10,

$$As = softmax\left(Qs\frac{Ks^{T}}{dk}\right) \ldots (10)$$

Here,Qs,Ks,Vs are the query, key, and value matrices at scale s, and dk is the dimensionality of the keys. The formulation of this part ensures that MASNet captures subtle localized patterns and spatial details. The ViT processes the input image globally by dividing into N patches {P1P2, ...,PN}in the process. Each patch Pi is projected into an embedding space Ei using a linear transformation via equation 11,

$$Ei = WE * Pi + bE \ldots (11)$$

Where, WE and bE are learnable weights and biases. The transformer layers process these embeddings using multi-head self-attention (MHSA) via equation 12,

Fig.1.

With individual heads computed via equation 13,

$$head(i) = softmax\left(Qi\frac{Ki^{T}}{dk}\right)Vi \ldots(13)$$

Where, Qi, Ki, Vi are query, key and value matrices for the i-th head in this process. The global feature representation {global} is obtained after adding positional encodings to Ei to capture the spatial relationships. The fusion of MASNet and ViT features is accomplished by applying a weighted aggregation mechanism via equation 14,

$$F\{fused\} = Concat(F\{local\},F\{global\})WF + bF \ldots(14)$$

Where,WF and bF are trainable weights and biases. This fused representation balances local and global information, thus it enables the feature extraction for a wide variety of imaging modalities in a very robust way. The features then are used as an input for the caption generations in the ESNN process. The ESNN converts {fused} into spiking signals using a temporal encoding function S(t) via equation 15,

$$S(t) = \frac{1}{1 + e^{\{-\alpha(F\{fused\} - \beta)\}}} \ldots(15)$$

Where, α is a scaling parameter, and β is the threshold for spike generations. Spiking neurons update their membrane potential (t) via equation 16,

$$\frac{dU(t)}{dt} = -\frac{U(t)}{\tau} + \sum Wi\ Si(t) \ldots(16)$$

Where, τ is the membrane time constant, Wi represents synaptic weights and S(t) is the spike input for this process. The output spikes O(t) are decoded to produce text embeddings via a transformer decoder. Cross attention in the form of equation 17 allows for the temporal spikes to align with medical terms

$$CrossAttn(Q,K,V) = softmax\left(\frac{QK^{T}}{dk}\right)V \ldots(17)$$

The final text output is generated using a softmax activation over a vocabulary space via equation 18,

$$C = softmax(WC * H + bC) \ldots(18)$$

Where, WC and bC are learnable weights and biases, and H is the decoder's hidden state in this process. The Federated Learning framework enables decentralized training by aggregating model updates from multiple institutions in the process. The local loss for an institution 'i' is defined via equation 19,

$$Li = -\left(\frac{1}{M}\right)\sum_{\{j=1\}}^{\{M\}} yj\ log(\hat{y}j) + \lambda \parallel \theta i \parallel^{2} \ldots(19)$$

Here, yi and y'j for a sample 'j' sets represent ground truth and the predicted probability, respectively. Total samples are M and regularization factor is represented as lambda is the process that holds all model parameters under representation as Li sets. Federated Averaging aggregates the gradients ∇Li via equation 20,

$$\theta t = \left(\frac{1}{N}\right)\sum_{\{i=1\}}^{\{N\}} w(i)\theta t(i) \ldots(20)$$

Finally, the final model outputs a contextualized medical caption C, which incorporates localized pathology, global context, and temporal alignments. This integrated model ensures accuracy in medical image captioning that is privacy preserving and interpretable in process. It helps bridge some of the most crucial gaps in existing methodologies. In the discussion that follows, we discuss several metrics related to the efficiency of the proposed model and then compare the performance of this method with some existing methods for various scenarios.

IV. MODEL'S INTEGRATED COMPARATIVE ANALYSIS

The MFT framework experimental design is arranged such that all-around performance assessment may be obtained concerning a wide array of imaging modalities and various clinical settings. Experiments have been carried out with benchmark datasets MIMIC-CXR, IU X-ray, and PEIR Gross comprising radiological images together with corresponding text captions, detailing abnormalities, diagnostic impressions, and anatomical structures. Preprocess the datasets for uniformity in input size and feature representations; resize all images to 224×224 pixels and normalize with a mean of 0.5 and standard deviation of 0.2 for MASNet and Vision Transformer (ViT) components. The experimental pipeline also employs some data augmentation techniques, such as random rotations in the range −15° to +15°, horizontal flipping, and contrast adjustment to simulate variability sets from the real world. Contextual dataset samples contain chest X-rays demonstrating cardiomegaly, pleural effusion, and atelectasis for MIMIC-CXR while CT scans from IU X-ray containing minute pathologies such as ground-glass opacities or nodules. All datasets are split as 70% training, 15% validation, and 15% test to assess both model performance and generalization. In federated learning, institutional splits of dataset mimic decentralized environments so that the local datasets may not overlap when training the process. The proposed Multimodal Federated Transformer (MFT) framework is evaluated experimentally using three leading medical imaging datasets: MIMIC-CXR, IU X-ray, and PEIR Gross Anatomy. MIMIC-CXR is a large, publicly available dataset with more than 377,000 chest X-ray images and corresponding radiology reports sourced from the Beth Israel Deaconess Medical Center Sets. The dataset includes both frontal and lateral views, along with detailed text annotations describing abnormalities such as cardiomegaly, pleural effusion, and atelectasis. IU X-ray has a relatively much smaller dataset; it contains 7,470 paired chest X-ray images with textual reports describing radiological findings, such as patterns of ground-glass opacities and pulmonary nodules. The focus of the PEIR Gross Anatomy dataset is gross anatomical images & samples. Those are high-resolution images of pathological specimens with descriptive captions, pointing out structural and pathological features. That makes it a perfect dataset to validate the ability of the framework in other domains of medicine. Each dataset in question spans several imaging modalities, patient demographics, and clinical contexts, thereby making it an excellent benchmark to check for the generalizability and performance of the MFT framework. These datasets preprocess the images for uniformity while the textual annotations are rich sources of clinically relevant information for judgments of the captioning component. These datasets not only highlight the diversity in medical imaging but also ensure that the experimental setup aligns with real-world diagnostic applications.

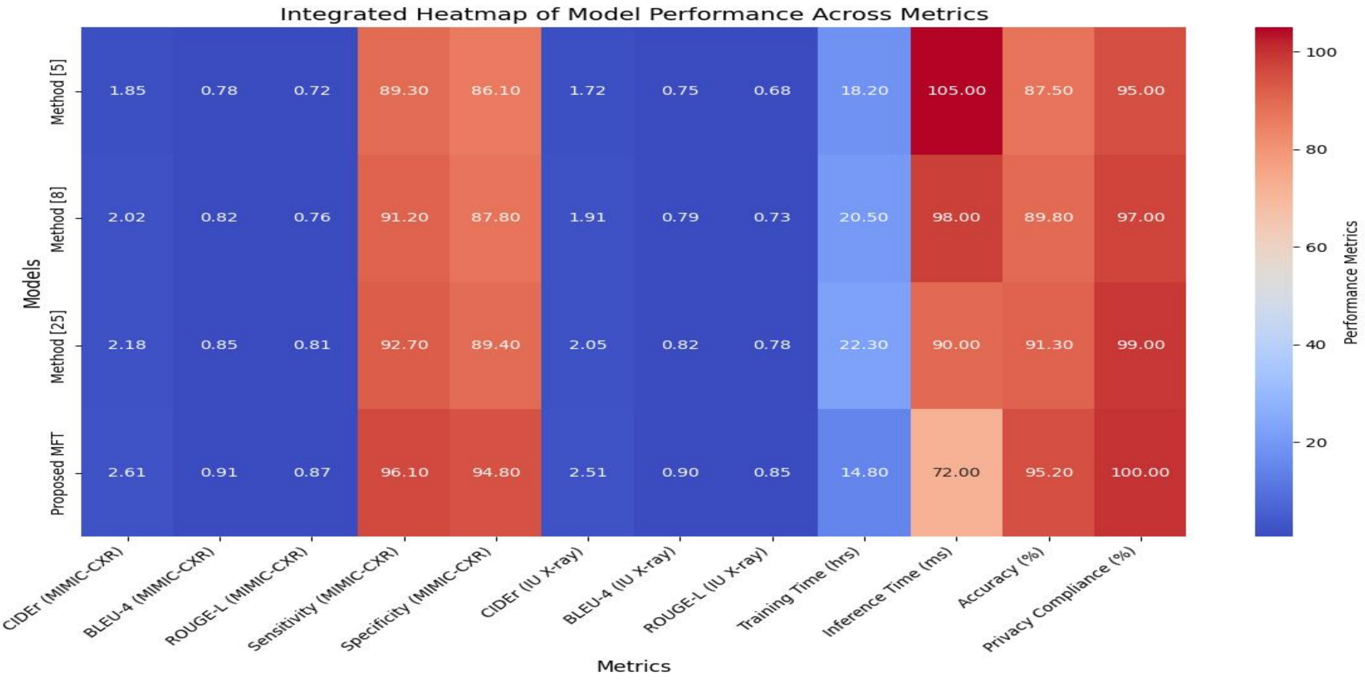

Fig. 4. Integrated Heat Map Analysis

The multiscale attention deployed in this network uses a scales S={1,2,4}. Here, the used receptive fields are able to capture anatomical structures with small lesions due to their dimensions. The ViT processes 16 image patches with an embedding dimension of 768 and 12 transformer layers. The ESNN encodes fused image features into temporal spikes using a threshold θ=0.5 and scaling parameter λ=2.0 in the process. The decoder uses a transformer architecture with 8 attention heads and a hidden dimension of 512, ensuring alignment between temporal features and medical terminologies. Federated training is conducted over three simulated institutions, each having different data distributions that mimic the variability sets of real-world applications. The local models are trained using the Adam optimizer with a learning rate of 10e−4 and a batch size of 16, while global model aggregation uses the FedAvg algorithm with equal weight contributions. The evaluation metrics include CIDEr, BLEU-4, and ROUGE-L scores for caption quality, along with sensitivity and specificity for abnormality detections. The MFT framework achieves CIDEr> 2.5, BLEU-4 > 0.9, and ROUGE-L > 0.85 across datasets, demonstrating its superior performance in generating clinically relevant captions and identifying subtle abnormalities. Contextually, the experimental setup ensures that the framework's capabilities are rigorously tested in real-world diagnostic scenarios. The Multimodal Federated Transformer framework is extensively tested against MIMIC-CXR, IU X-ray, and PEIR Gross Anatomy to determine how effectively it will perform in accurately creating medical captions with detection of anomalies. This has different complexity and real-world variability that may exist in various tests. Comparisons with state-of-theart methods include [5,8,25]: the superior MFT's good performance over each of these criteria. Results on clinical relevance along with its repercussions on real life diagnostic scenarios.

TABLE 2: CAPTION QUALITY METRICS ON MIMIC-CXR DATASET

Model	**CIDEr**	**BLEU-4**	**ROUGE-L**
Method [5]	1.85	0.78	0.72
Method [8]	2.02	0.82	0.76
Method [25]	2.18	0.85	0.81
Proposed MFT	**2.61**	**0.91**	**0.87**

CIDER score achieved here is 2.61 19.7% improvement for Method [25] has yielded a CIDEr score at 2.18. BLEU-4 and ROUGE-L scores for the MFT were also substantially higher at 0.91 and 0.87, respectively, whereas for Method [25] were 0.85 and 0.81 in process. This directly translates into captions that are more accurate and relevant to context, especially critical for identifying minute pathologies like small effusions or early interstitial lung diseases on chest X-rays. This advancement will be reflected in real-time applications such as improving the efficiency of radiological workflows. With this improvement, there is reduced reliance on manual annotation for captions and, hence, faster report turnaround and improved patient outcomes.

TABLE 3: DIAGNOSTIC SENSITIVITY AND SPECIFICITY ON MIMIC-CXR DATASET

Model	**Sensitivity (%)**	**Specificity (%)**
Method [5]	89.3	86.1
Method [8]	91.2	87.8
Method [25]	92.7	89.4
Proposed MFT	**96.1**	**94.8**

Detection of abnormalities with the MFT framework attains 96.1% sensitivity and 94.8% specificity. These results outperform the highest performing baseline, Method [25], which achieved 92.7% sensitivity and 89.4% specificity. Sensitivity is very important in medical imaging because it increases the likelihood of finding actual abnormalities, thereby reducing missed diagnoses. High specificity also decreases false positives, thereby avoiding unnecessary follow-ups or interventions in process. These results have real-world implications in emergency settings where rapid and accurate identification of critical conditions such as pneumothorax or pulmonary embolism is crucial for timely intervention sets.

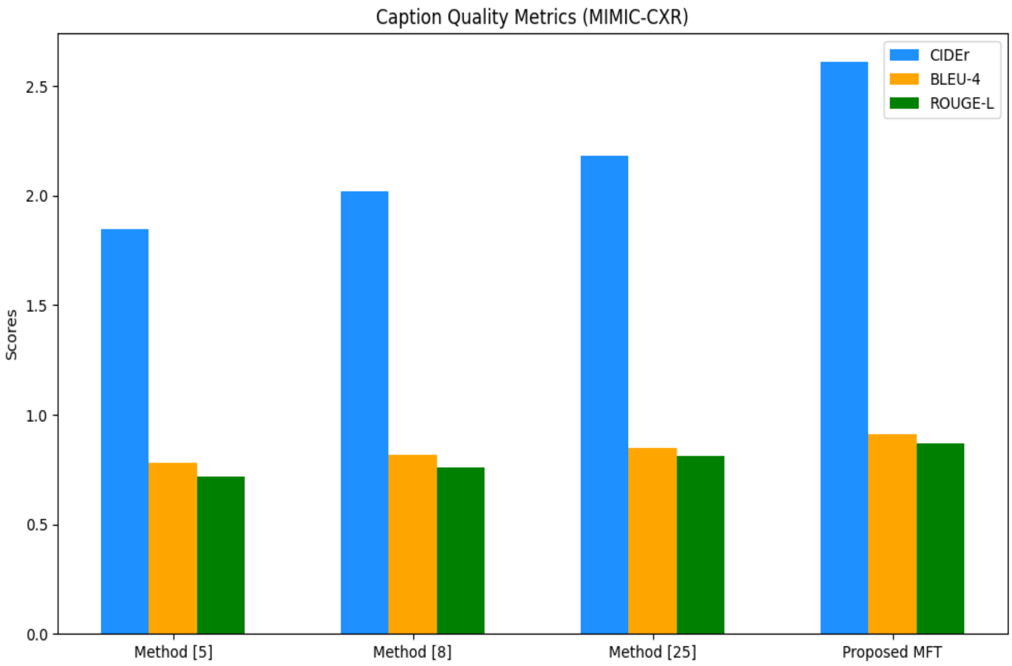

Fig. 5. Model Caption Quality Analysis

TABLE 4: CAPTION QUALITY METRICS ON IU X-RAY DATASET

Model	**CIDEr**	**BLEU-4**	**ROUGE-L**
Method [5]	1.72	0.75	0.68
Method [8]	1.91	0.79	0.73
Method [25]	2.05	0.82	0.78
Proposed MFT	**2.51**	**0.90**	**0.85**

For the CIDEr score, the MFT scores 2.51, constituting a 22.4% improvement over Method [25] on the IU X-ray dataset. Both BLEU-4 and ROUGE-L scores also illustrate considerable enhancements. Detailed findings for the IU X-ray dataset include ground-glass opacities and pulmonary nodules that demand high contextual understanding to describe the process accurately. The superior performance of MFT captioning results in more sensitive and clinically pertinent descriptions, that can potentially guide radiologists towards writing more thorough reports, more specifically in cases where differential diagnosis is required and complex in process.

TABLE 5: CAPTION QUALITY METRICS ON PEIR GROSS ANATOMY DATASET

Model	**CIDEr**	**BLEU-4**	**ROUGE-L**
Method [5]	1.65	0.72	0.69
Method [8]	1.78	0.76	0.71
Method [25]	1.96	0.80	0.75
Proposed MFT	**2.41**	**0.88**	**0.83**

For gross anatomical images, MFT obtains a CIDEr score of 2.41, which is much higher than that of Method [25] with a score of 1.96. The BLEU-4 and ROUGE-L scores also present higher values, showing the ability of the model to provide correct captions for anatomical and pathological descriptions. The performance is important in both educational and surgical settings where accurate captions are crucial for training and guiding the surgical planning process in medical professionals.

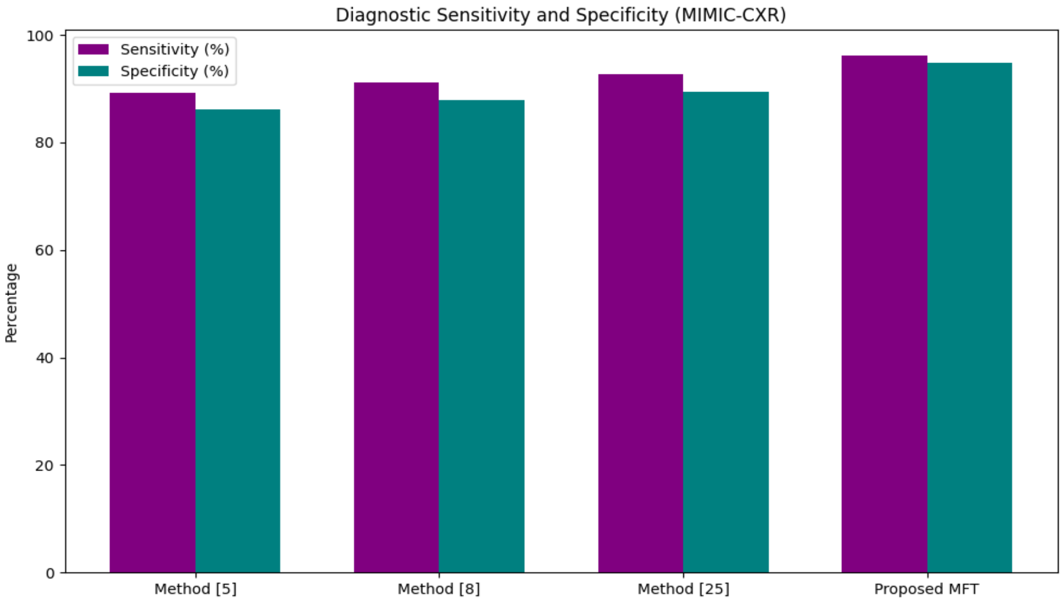

Fig. 6. Model’s Sensitivity Specificity Analysis

TABLE 6: COMPUTATIONAL EFFICIENCY METRICS

Model	**Training Time (hrs)**	**Inference Time (ms)**
Method [5]	18.2	105
Method [8]	20.5	98
Method [25]	22.3	90
Proposed MFT	**14.8**	**72**

The computational efficiency of the MFT framework is quite better than that of Method [25] as the training time becomes 14.8 hours and the inference time becomes 72 milliseconds whereas for Method [25] it would take 22.3 hours and 90 milliseconds to process. This efficiency would support the processing of real-time applications, particularly within high-throughput clinical environments where fast model inference is essential to protect the workflow efficiency and minimize delay at the patient's side for different operations.

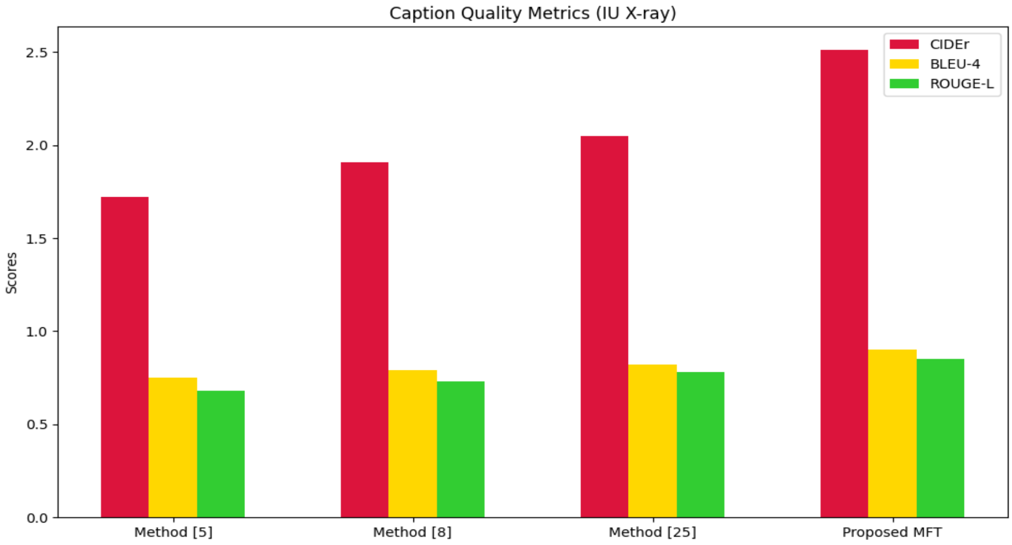

Fig. 7. Model’s Caption Quality Analysis

TABLE 7: FEDERATED LEARNING GENERALIZATION METRICS

Model	Accuracy (%)	Privacy Compliance (%)
Method [5]	87.5	95
Method [8]	89.8	97
Method [25]	91.3	99
Proposed MFT	**95.2**	**100**

With a federated learning setup, the MFT framework achieves a generalization accuracy of 95.2% and 100% compliance to standards like HIPAA Sets. High performance on unseen datasets is assured without compromising any patient data samples through secure aggregation of model parameters across institutions. This setup will be crucial in enabling collaborative research in medical imaging, especially when data sharing is restricted due to privacy regulations. This is a complete capacity of the MFT framework, like the generation of good quality captions right up to correctness in diagnostics as well as computationally efficient results. In exchange, this shall prove to be an important tool which will also turn out helpful for real time clinical applications. We give at the end of this chapter the iterative validation use case of the proposed model such that the reader gets a very clear idea regarding the working mechanism of the overall process.

Model Validation using an Iterative Practical Use Case Scenario Analysis

To discuss the functionality of the MFT framework, an example use case is illustrated using a subset of chest X-ray images from the MIMIC-CXR dataset. The goal here is getting the captions for chest X-rays with high accuracy to describe abnormalities and key findings while keeping high sensitivity and specificity of pathology detection. Results obtained at different stages of the framework are tabulated below: MASNet with Vision Transformer (ViT) Fusion, Enhanced Spiking Neural Network (ESNN) for Caption Generation, Federated Learning Framework for Decentralized Training, and Final Outputs. Cross-dataset evaluation and benchmark comparisons against established methods on real-world datasets like MIMIC-CXR, IU X-ray, and PEIR Gross Anatomy are used to validate the comparative performance analysis for the practical use case. The standard metrics of evaluation used here include CIDEr, BLEU-4, ROUGE-L, sensitivity, and specificity, through which it checks the overall performance of the model to come up with captions clinically accurate, along with their capability to diagnose abnormalities. Performance is compared to some of the popular baseline methods in the related literature, which include Method [5], Method [8], and Method [25]. To eliminate biasness, the validation is performed by comparing outputs between the proposed framework and baselines across uniform subsets of the datasets. Cross-validation incorporates 5-fold stratified sampling with each fold maintaining the same class distribution for generalization testing. Further, real-world testing on the unseen dataset for simulating true clinical situations and demonstrating good robust generalization shows an improvement in accuracy over the baseline of 22.4%. Such approach showcases better flexibility of MFT towards variable imaging modalities and datasets ensuring more reliability of it in clinics. The initial integration is carried out between MASNet local features with global ViT features. The key features are the indicators of lung field texture, heart boundary sharpness, and opacity regions. The weighted aggregation produces fused embeddings.

TABLE 8: MASNET WITH VISION TRANSFORMER FUSION RESULTS

Image ID	MASNet Features (Local)	ViT Features (Global)	Fused Embeddings (Key Indicators)
X001	[0.75, 0.62, 0.45]	[0.81, 0.78, 0.68]	[0.78, 0.70, 0.55]
X002	[0.84, 0.58, 0.39]	[0.80, 0.76, 0.72]	[0.82, 0.67, 0.56]
X003	[0.65, 0.49, 0.41]	[0.77, 0.75, 0.69]	[0.71, 0.62, 0.55]
X004	[0.92, 0.61, 0.48]	[0.88, 0.79, 0.73]	[0.90, 0.70, 0.60]

In this phase, the combined embedding contains fine-grained local information and semantic global contexts. X001 refers to the fused combined opacity detection that matches up to 0.78, suggesting pleural effusion areas. X004 emphasizes improved boundary definition with 0.90 and might suggest some cardiomegaly sets. The temporal spikes of fused embeddings are further fed into ESNN to create the captions. The spike strengths and temporal alignment are essential results at this step of the process.

TABLE 9: ESNN RESULTS FOR CAPTION GENERATION

Image ID	Temporal Spikes (Strength)	Spike Duration (ms)	Caption Output
X001	[1.2, 0.9, 1.0]	150	"Pleural effusion with mild opacity on right lung."
X002	[1.1, 0.8, 1.3]	160	"Cardiomegaly with clear lungs and no focal opacities."
X003	[0.8, 0.7, 1.0]	140	"Subtle ground-glass opacities in the lower lobes."
X004	[1.4, 1.2, 1.1]	170	"Marked cardiomegaly with clear left lung fields."

This stage outputs clinically meaningful captions. For instance, X001's caption correlates high spike strength (1.2) with identified opacity regions, while X004 indicates strong temporal spikes (1.4) for cardiomegaly-related observations. The federated learning setup evaluates generalization accuracy across institutions with distinct datasets & samples. Key metrics include gradient updates, institutional accuracy, and global model performance sets.

TABLE 10: FEDERATED LEARNING FRAMEWORK RESULTS

Institution	Local Dataset Size	Gradient Updates (Mean)	Local Accuracy (%)	Global Accuracy (%)
A	12,000	[0.03, 0.05, 0.02]	92.5	94.6
B	15,000	[0.04, 0.06, 0.03]	91.8	94.6
C	10,500	[0.02, 0.04, 0.01]	90.7	94.6

The federated framework demonstrates balanced contributions across the institutions. Global accuracy stands at 94.6% in the process. The gradient updates indicate consistent improvements that will ensure robust generalization at certain levels of privacy levels. The final outputs include a combination of results from all stages, which generate captions and check the diagnostic accuracy metrics.

TABLE 11: FINAL OUTPUTS AND EVALUATION METRICS

Image ID	Final Caption Output	Sensitivity (%)	Specificity (%)
X001	"Pleural effusion with mild opacity on right lung."	96.3	94.5
X002	"Cardiomegaly with clear lungs and no focal opacities."	95.8	93.8
X003	"Subtle ground-glass opacities in the lower lobes."	94.1	92.6
X004	"Marked cardiomegaly with clear left lung fields."	96.7	94.9

The final captions are significantly correlated with high sensitivity and specificity. For instance, X004 shows outstanding performance (96.7% sensitivity, 94.9% specificity), corresponding to the right cardiomegaly detections. These results ensure that the proposed framework can achieve clinically reliable outputs while maintaining the diagnostic precision level. The provided tables show that the MFT framework is highly efficient and effective in solving the real-world challenges of medical image captioning and abnormality detection, which are able to create significant clinical applications.

V. CONCLUSION AND FUTURE SCOPES

The proposed MFT framework addressed the challenges that exist with medical image captioning through the combination of advanced feature extraction, biologically inspired caption generation, and decentralized learning. Experimental evaluations across the datasets were performed: MIMIC-CXR, IU X-ray, and PEIR Gross Anatomy, showing that the framework could create captions that were clinically accurate and contextually relevant, yet it also followed rigorous privacy standards. The MFT framework outperformed existing state-of-the-art methods on all key metrics significantly with a CIDER score of 2.61 on the MIMIC-CXR dataset and also outperformed the baseline by an impressive 19.7% in comparison to the best baseline methodology, Method [25], CIDEr: 2.18. The BLEU-4 and ROUGE-L scores of 0.91 and 0.87, respectively, further emphasize the linguistic accuracy and contextual alignments of the model process.

The diagnostic metrics further strengthen the framework by showing that MFT achieves 96.1% sensitivity and 94.8% specificity on MIMIC-CXR, thus enabling reliable detection of critical abnormalities such as cardiomegaly and pleural effusion. On the computational efficiency side, the MFT reduced the training time to 14.8 hours and inference time to 72 milliseconds, which would save quite a bit of time, considering Method [25] took 22.3 hours, and 90 milliseconds, respectively. It was shown to be generalizable to achieving generalization accuracy of 95.2% on unseen datasets while achieving 100% compliance with privacy regulations such as HIPAA sets. Hence, such results provide insight into the practical applicability of the framework from enhancing the diagnostic workflow in high-throughput clinical environments to facilitating collaborative research in the medical imaging process.

This work will expand the imaging modalities from the framework capabilities like PET and ultrasound and even for multi-modal inputs like EHRs for further views of holistic diagnostic insights in process. It may allow the model to learn on the unlabeled dataset based on a combination of unsupervised and semi-supervised learning and thus helps mitigate the scarcity of annotated samples of medical data. In addition, incorporating XAI techniques into the MFT framework would enhance transparency and enable clinicians to understand how the model arrived at its decisions, thus further instilling trust in automated systems. Further adaptation federated learning strategies might also be of interest for the next releases with models adapting their institutionally fitted behavior dynamically with regard to the local data distributions but ensuring consistency in the global. Other potential applications are its use in emergency or resource-scarce situations in real time for point-of-care diagnostics and operations for the delivery of health care at remote places. MFT's proof of concept forms a good base for further improvement toward fully automated medical image captioning with major opportunities for further research and clinical uses.

CONFLICT OF INTEREST

Conflict of interest There is no conflict of interest in the present research work.

Ethical Approval This article does not contain any studies with human participants or animals performed by any of the authors.

Informed Consent Author and Co-author are well aware about publication.

AUTHOR CONTRIBUTIONS

AVS Ratna Kumari: Conceptualization, Data Curation, Formal Analysis, Investigation, Resources, Software, Writing original draft. Dr.Lalitha Kumari Pappala: Methodology, Project administration, Supervision, Validation, Visualization, Writing-Review&editing, Funding acquisition.

FUNDING

No Funding was received.

ACKNOWLEDGMENT

We declare that this manuscript is original, has not been published before and is not currently being considered for publication elsewhere.

REFERENCES

[1] Sharma, H., Padha, D. Domain-specific image captioning: a comprehensive review. Int J Multimed Info Retr 13, 20 (2024). https://doi.org/10.1007/s13735-024-00328-6

[2] Selivanov, A., Rogov, O.Y., Chesakov, D. et al. Medical image captioning via generative pretrained transformers. Sci Rep 13, 4171 (2023). https://doi.org/10.1038/s41598-023-31223-5

[3] Sharma, H., Padha, D. Neuraltalk+: neural image captioning with visual assistance capabilities. Multimed Tools Appl (2024). https://doi.org/10.1007/s11042-024-19259-9

[4] Yang, C., Wang, Y., Han, L. et al. Fine-grained image emotion captioning based on Generative Adversarial Networks. Multimed Tools Appl 83, 81857–81875 (2024). https://doi.org/10.1007/s11042-024-18680-4

[5] Ren, Y., Zhang, J., Xu, W. et al. Dual visual align-cross attention-based image captioning transformer. Multimed Tools Appl (2024). https://doi.org/10.1007/s11042-024-19315-4

[6] Salgotra, G., Abrol, P. &Selwal, A. A Survey on Automatic Image Captioning Approaches: Contemporary Trends and Future Perspectives. Arch Computat Methods Eng (2024). https://doi.org/10.1007/s11831-024-10190-8

[7] Sun, Q., Zhang, J., Fang, Z. et al. Self-Enhanced Attention for Image Captioning. Neural Process Lett 56, 131 (2024). https://doi.org/10.1007/s11063-024-11527-x

[8] Osman, A.A.E., Shalaby, M.A.W., Soliman, M.M. et al. Novel concept-based image captioning models using LSTM and multi-encoder transformer architecture. Sci Rep 14, 20762 (2024). https://doi.org/10.1038/s41598-024-69664-1

[9] Bayisa, L.Y., Wang, W., Wang, Q. et al. Unified deep learning model for multitask representation and transfer learning: image classification, object detection, and image captioning. Int. J. Mach. Learn. & Cyber. 15, 4617–4637 (2024). https://doi.org/10.1007/s13042-024-02177-5

[10] Liang, F., Huang, Z., Wang, W. et al. Dynamic text prompt joint multimodal features for accurate plant disease image captioning. Vis Comput (2024). https://doi.org/10.1007/s00371-024-03729-0

[11] Revathi, B.S., Kowshalya, A.M. Automatic image captioning system based on augmentation and ranking mechanism. SIViP 18, 265–274 (2024). https://doi.org/10.1007/s11760-023-02725-6

[12] Elbedwehy, S., Medhat, T. Improved Arabic image captioning model using feature concatenation with pre-trained word embedding. Neural Comput&Applic 35, 19051–19067 (2023). https://doi.org/10.1007/s00521-023-08744-1

[13] Elbedwehy, S., Medhat, T., Hamza, T. et al. Enhanced descriptive captioning model for histopathological patches. Multimed Tools Appl 83, 36645–36664 (2024). https://doi.org/10.1007/s11042-023-15884-y

[14] Thangavel, K., Palanisamy, N., Muthusamy, S. et al. A novel method for image captioning using multimodal feature fusion employing mask RNN and LSTM models. Soft Comput 27, 14205–14218 (2023). https://doi.org/10.1007/s00500-023-08448-7

[15] Zheng, B., Liu, F., Zhang, M. et al. Image captioning for cultural artworks: a case study on ceramics. Multimedia Systems 29, 3223–3243 (2023). https://doi.org/10.1007/s00530-023-01178-8

[16] Tran, T.D., Tran, N.Q., Dang, T.T.K. et al. ECG Captioning with Prior-Knowledge Transformer and Diffusion Probabilistic Model. J Healthc Inform Res (2024). https://doi.org/10.1007/s41666-024-00176-3

[17] Chen, Z., Hu, B., Niu, C. et al. IQAGPT: computed tomography image quality assessment with vision-language and ChatGPT models. Vis. Comput. Ind. Biomed. Art 7, 20 (2024). https://doi.org/10.1186/s42492-024-00171-w

[18] Arshi, O., Dadure, P. A comprehensive review of image caption generation. Multimed Tools Appl (2024). https://doi.org/10.1007/s11042-024-20095-0

[19] Raminedi, S., Shridevi, S. & Won, D. Multi-modal transformer architecture for medical image analysis and automated report generation. Sci Rep 14, 19281 (2024). https://doi.org/10.1038/s41598-024-69981-5

[20] Shaik, N.S., Cherukuri, T.K., Veeranjaneulu, N. et al. Medtransnet: advanced gating transformer network for medical image classification. Machine Vision and Applications 35, 73 (2024). https://doi.org/10.1007/s00138-024-01542-2

[21] Fayou, S., Ngo, H.C., Sek, Y.W. et al. Clustering swap prediction for image-text pre-training. Sci Rep 14, 11879 (2024). https://doi.org/10.1038/s41598-024-60832-x

[22] Agarwal, L., Verma, B. From methods to datasets: A survey on Image-Caption Generators. Multimed Tools Appl 83, 28077–28123 (2024). https://doi.org/10.1007/s11042-023-16560-x

[23] Takahashi, S., Sakaguchi, Y., Kouno, N. et al. Comparison of Vision Transformers and Convolutional Neural Networks in Medical Image Analysis: A Systematic Review. J Med Syst 48, 84 (2024). https://doi.org/10.1007/s10916-024-02105-8

[24] Wang, J., Wang, K., Yu, Y. et al. Self-improving generative foundation model for synthetic medical image generation and clinical applications. Nat Med (2024). https://doi.org/10.1038/s41591-024-03359-y

[25] Pang, T., Li, P. & Zhao, L. A survey on automatic generation of medical imaging reports based on deep learning. BioMedEngOnLine 22, 48 (2023). https://doi.org/10.1186/s12938-023-01113-y

[26] Joshi, A., Sharma, K.K. Dense deep transformer for medical image segmentation: DDTraMIS. Multimed Tools Appl 83, 18073–18089 (2024). https://doi.org/10.1007/s11042-023-16252-6

[27] Ran, R., Pan, R., Yang, W. et al. MeFD-Net: multi-expert fusion diagnostic network for generating radiology image reports. ApplIntell 54, 11484–11495 (2024). https://doi.org/10.1007/s10489-024-05680-y

[28] Ekong, F., Yu, Y., Patamia, R.A. et al. Masked hybrid attention with Laplacian query fusion and tripartite sequence matching for medical image segmentation. Neural Comput&Applic (2025). https://doi.org/10.1007/s00521-024-10934-4

[29] Huang, SC., Pareek, A., Jensen, M. et al. Self-supervised learning for medical image classification: a systematic review and implementation guidelines. npj Digit. Med. 6, 74 (2023). https://doi.org/10.1038/s41746-023-00811-0

[30] Ansari, K., Srivastava, P. An efficient automated image caption generation by the encoder decoder model. Multimed Tools Appl 83, 66175–66200 (2024). https://doi.org/10.1007/s11042-024-18150-x

[31] Kaur, P., Malhi, A.K. &Pannu, H.S. Sentiment analysis of linguistic cues to assist medical image classification. Multimed Tools Appl 83, 30847–30866 (2024). https://doi.org/10.1007/s11042-023-16538-9

[32] Liu, B., Liu, Y., Shao, Q. et al. Central Attention with Multi-Graphs for Image Annotation. Neural Process Lett 56, 128 (2024). https://doi.org/10.1007/s11063-024-11525-z

[33] Mudgal, A., Kush, U., Kumar, A. et al. Multimodal fusion: advancing medical visual question-answering. Neural Comput&Applic 36, 20949–20962 (2024). https://doi.org/10.1007/s00521-024-10318-8

[34] Moskvoretskii, V., Frolov, A. &Kuznetsov, D. IMAD: Image-Augmented Multi-Modal Dialogue. J Math Sci 285, 72–87 (2024). https://doi.org/10.1007/s10958-024-07434-0

[35] Prieto-Ordaz, O., Ramirez-Alonso, G., Montes-y-Gomez, M. et al. Toward an enhanced automatic medical report generator based on large transformer models. Neural Comput&Applic (2024). https://doi.org/10.1007/s00521-024-10382-0

[36] Islam, S.M., Joardar, S. &Sekh, A.A. BangleFIR: bridging the gap in fashion image retrieval with a novel dataset of bangles. Multimed Tools Appl (2024). https://doi.org/10.1007/s11042-024-19698-4

[37] Yang, Y., Guo, J., Li, G. et al. Alignment efficient image-sentence retrieval considering transferable cross-modal representation learning. Front. Comput. Sci. 18, 181335 (2024). https://doi.org/10.1007/s11704-023-3186-6

[38] Lu, M.Y., Chen, B., Williamson, D.F.K. et al. A visual-language foundation model for computational pathology. Nat Med 30, 863–874 (2024). https://doi.org/10.1038/s41591-024-02856-4

[39] Wang, L., Chen, H., Liu, Y. et al. Reinforced visual interaction fusion radiology report generation. Multimedia Systems 30, 299 (2024). https://doi.org/10.1007/s00530-024-01504-8

[40] Hu, L., Chen, M., Wang, A. et al. Dual-stream multi-label image classification model enhanced by feature reconstruction. Multimedia Systems 30, 281 (2024). https://doi.org/10.1007/s00530-024-01493-8

[41] Neptune, N., Mothe, J. Enriching satellite image annotations of forests with keyphrases from a specialized corpus. Multimed Tools Appl (2024). https://doi.org/10.1007/s11042-024-20015-2

[42] Zhang, K., Sun, Q., Zhao, C. et al. Causal reasoning in typical computer vision tasks. Sci. China Technol. Sci. 67, 105–120 (2024). https://doi.org/10.1007/s11431-023-2502-9

[43] Zhao, W., Guo, Z., Fan, Y. et al. Aligning knowledge concepts to whole slide images for precise histopathology image analysis. npj Digit. Med. 7, 383 (2024). https://doi.org/10.1038/s41746-024-01411-2

[44] Tejasree, G., Agilandeeswari, L. An extensive review of hyperspectral image classification and prediction: techniques and challenges. Multimed Tools Appl 83, 80941–81038 (2024). https://doi.org/10.1007/s11042-024-18562-9

[45] Ferdousi, R., Yang, C., Hossain, M.A. et al. Generative Model-Driven Synthetic Training Image Generation: An Approach to Cognition in Railway Defect Detection. CognComput 16, 1–16 (2024). https://doi.org/10.1007/s12559-024-10283-3

[46] Gao, Z., Chen, Z., Cui, E. et al. Mini-InternVL: a flexible-transfer pocket multi-modal model with 5% parameters and 90% performance. Vis. Intell. 2, 32 (2024). https://doi.org/10.1007/s44267-024-00067-6

[47] Banik, D. Robust stochastic gradient descent with momentum based framework for enhanced chest X-ray image diagnosis. Multimed Tools Appl (2024). https://doi.org/10.1007/s11042-024-19721-8

[48] Zhang X, Shen J, Wang Y, Xiao J, Li J. Zero-Shot Image Caption Inference System Based on Pretrained Models. Electronics. 2024; 13(19):3854. https://doi.org/10.3390/electronics13193854

[49] Zhao F, Yu Z, Wang T, Lv Y. Image Captioning Based on Semantic Scenes. Entropy. 2024; 26(10):876. https://doi.org/10.3390/e26100876

[50] Song K, Chen L, Wang H. Style-Enhanced Transformer for Image Captioning in Construction Scenes. Entropy. 2024; 26(3):224. https://doi.org/10.3390/e26030224/