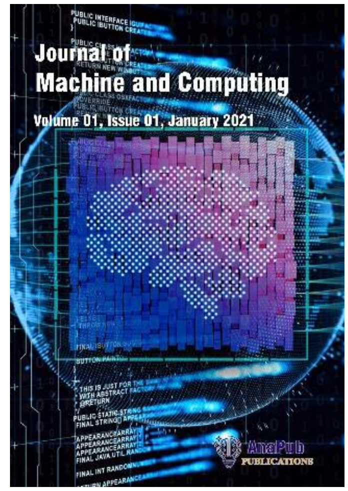
Journal Pre-proof



MetaFusion-FL: A Cross-Modality Federated Meta-Learning Framework for Robust and Explainable Healthcare System

Kalphana KR, Maheskumar V, Vijayarajeswari R and Sasikala K

DOI: 10.53759/7669/jmc202505141

Reference: JMC202505141

Journal: Journal of Machine and Computing.

Received 02 April 2025

Revised from 28 May 2025

Accepted 16 June 2025

AnaPub
PUBLICATIONS

MetaFusion-FL: A Cross-Modality Federated Meta-Learning Framework for Robust and Explainable Healthcare System

[1] KR.Kalphana*, [2] V.Maheskumar, R.Vijayarajeswari, K. Sasikala

[1]Associate Professor &Head, Department of Agricultural Engineering, Mahendra Engineering College, Namakkal, Tamil Nadu, India.

[2]Assistant Professor, Department of Computer Science and Engineering, Paavai Engineering College, Namakkal, Tamil Nadu, India.

[3]Associate Professor, Department of Information Technology (AI&DS), Sona College of Technology, Salem, Tamil Nadu, India.

[4]Professor, Department of Information Technology, R P Sarathy Institute of Technology, Salem, Tamil Nadu, India.

hodagri@mahendra.info, mahestamil@gmail.com, vijimecse@gmail.com, email2sasi@gmail.com

*Corresponding Author: **KR.Kalphana**

Abstract

Mpox is a re-emerging zoonotic viral disease that attracted the attention of the whole world because of its spreading transmission and clinical similarity with other skin diseases. It is highly important that this identification is fast and accurate, even in remotely located areas or resource-limited settings. However, the conventional centralized deep learning models exhibit severe limitations regarding data privacy, modality variation, and scalability across varied clinical environments. To this end, this paper presents MetaFusion-FL, a new federated meta-learning framework that combines cross-modality image analysis based on a hybrid Transformer-Capsule model with Hierarchical Attention-Based Multimodal Fusion (HAMFM). The model can work on multi-source images as input, namely smartphone images, dermoscopic images, and clinical images, which are processed locally at edge hospitals without raw data transmission. Reptile federated meta-learning strategy guarantees quick personalization of models and global generalization. When evaluated on a wide dataset, MetaFusion-FL has a higher classification accuracy of 99.46%, precision of 99.52%, recall of 99.40%, and F1-score of 99.46% compared to other current models, including ViT-RLXGBFL (99.12%) and ResViT-FLBoost (98.78%). The framework is also resistant to image noise and is consistent and stable across federated clients. Besides, SHAP and Grad-CAM++ explanations are used to ensure interpretability in a clinical context. MetaFusion-FL is therefore a leap in the development of AI-based, privacy-preserving, and generalizable skin disease classification, particularly Mpox.

Keywords: Mpox Detection, Cross-Modality, Federated Learning, Meta-Learning, Capsule Network, Transformer, Medical Imaging, Multimodal Fusion.

1. Introduction

Mpox (Monkeypox) is a viral zoonotic infection that has currently acquired global consideration after its re-emergence and the possibility of human-to-human spreading. Historically, the disease has been endemic to Central and West Africa, but recent outbreaks have occurred across the world, leading to questions of how quickly the disease can be diagnosed and contained [1] [2]. The diagnosis Early and proper diagnosis plays an important role during the management of an outbreak, decreases transmission, and provides proper clinical care. Although the traditional diagnostic techniques like PCR (polymerase chain reaction) and ELISA (enzyme-linked immunosorbent assay) are sensitive, they are time consuming, expensive, and need special laboratory conditions. The latter has spurred the desire to use artificial intelligence, in particular deep learning models, to perform swift, non-invasive Mpox detection based on images of skin lesions [3] [4].With regard to medical image classification, deep learning, namely convolutional neural networks (CNNs) and transformer-based models, including ViT (Vision Transformer) and SwinTransformer have shown promise. Such models in Mpox could be used to process high-resolution images of skin lesions and differentiate between Mpox and other skin diseases that can exhibit similarly, including chickenpox or syphilis. The benefit of such models will be the presence of pattern recognition and possible decision support in real-time, both in the clinical and remote setting [5] [6]. In addition to that, the use of self-supervised learning and data augmentation techniques helps to increase the resilience of the models despite the limited size of the annotated dataset, and in case of an outbreak, when the speed of implementation is a priority, deep learning is an attractive suggestion [7] [8].

However, despite these promising results, there are several limitations to the use of deep learning in the detection of Mpox. One is the challenge of availability of massive, diverse, and quality annotated databanks. The datasets used are mostly geographically or demographically limited which may reduce the applicability of the models to the populations of the world. Second, Mpox skin lesions may resemble other skin diseases, and with inadequately trained model, false positive or negative outcome will be obtained [9] [10] [11]. Third, most models would be deemed black boxes, i.e. there would be minimal interpretation of the prediction which could hinder clinical trust and adoption.

The significant downside of deep learning models, in terms of Mpox detection, is the necessity in data quality and quantity. The inconsistency of the lighting systems, quality of the cameras, or resolution of the images can produce a significant effect on the accuracy of the predictions. Additionally, those models are computation-demanding both in training and inference, which may not be possible in resource-limited settings where Mpox is most prevalent [12] [13]. There exists also the problem of algorithmic bias: models trained on one cohort of individuals may underperform on other skin colors, ages, or clinical contexts, threatening to increase healthcare inequities even more. Lastly, deep learning models have earned the unenviable reputation of being non-transparent, meaning that a clinician would not easily be able to answer why a prediction was made, which can inhibit their use in the clinical workflow [14] [15]. Although deep learning models present a recent and potentially revolutionary solution to the detection of Mpox, a couple of limitations and weaknesses need to be addressed. Dataset standardization, model interpretability, and fair model deployment remain active areas of research to make sure such technologies can be used ethically and effectively in global public health response. Figure 1 shows the symptoms of Mpox.

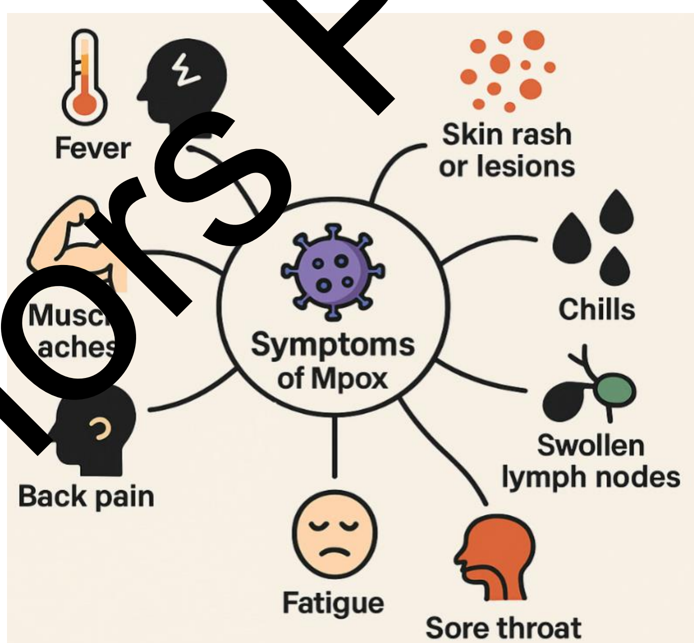


Figure 1: Symptoms of Mpox

In order to mitigate these shortcomings, this paper proposes MetaFusion-FL, a cross-modality federated meta-learning framework toward robust, accurate, and explainable Mpox detection. The model we propose brings together a few novelties: (1) we use Hierarchical Attention-Based Multimodal Fusion (HAMFM) to fuse features extracted from smartphone, dermoscopic, and clinical images; (2) we encode data using a hybrid Transformer-Capsule encoder to capture both long-range dependencies and morphological hierarchies in lesions; and (3) we use the Reptile federated meta-learning algorithm to guarantee fast adaptation and weight convergence across all clients without requiring data sharing. By combing these elements in a federated setting, each healthcare institution can train a local model locally, and contributes to a global model without sending sensitive patient information.

1.1. Main Contribution of the Work

- Cross-Modality Image Fusion Architecture:Proposed a new architecture that integrates dermoscopic, smartphone, and clinical imaging modalities with hierarchical attention mechanisms, which can reliably perform across image sources which are otherwise heterogeneous.
- Hierarchical Attention-Based Multimodal Fusion (HAMFM): Presented a new fusion block with channel-wise, spatial, and modality-aware attention mechanisms to highlight the features of the lesion area and preserve the modality attributes.
- Hybrid Transformer-Capsule Feature Encoder: Proposed an encoder layer to incorporate transformer blocks to represent global context and capsule network to part-to-whole lesion morphology to make the diagnosis more robust.
- Federated Meta-Learning in Reptile Optimization: Introduced a privacy-preserving federated learning procedure founded on the Reptile optimizer, allowing client-level personalization without the centralization of the data.
- Modality-Invariant Feature Weighting with XGBoost: Developed a modality-invariant feature importance enhancement mechanism XGBoost that enables final-stage classification and interpretability across modalities.
- Artifact-Robust Preprocessing Pipeline: Designed a standardized preprocessing pipeline, Image modality-wise, of CLAHE, adaptive thresholding, hybrid noise filtering, and Z-score normalization.

The rest of the paper is structured as follows. Section 2 thoroughly describes related works involving Mpox detection, federated learning on medical images, and multimodal fusion strategies, revealing the drawbacks of the current frameworks. Section 3 elaborates the proposed MetaFusion-FL methodology, explaining the cross-modality fusion pipeline, Transformer-Capsule encoding, hierarchical attention designs, and federated meta-learning plan. Section 4 reports the experimental findings, performance analysis and comparison with benchmark models. Lastly, Section 5 concludes the paper, summing up the essential findings and providing the prospect of the real-world implementation, extension to multiple diseases, and adapting to changing clinical conditions.

2.Related Work

In a narrative review, the association of Mpox virus (MPXV) infection and the diagnosing ability of saliva was noted. The MPXV replicated with the aid of endoplasmic reticulum, ribosomes as well as cytoplasmic proteins of the host cell. Lesions on the oral mucosa were frequent prior to skin rashes and the conventional diagnostic methods were unable to identify the virus early. A transmission medium, Saliva, was promising as a non-invasive diagnostic fluid [16]. In small-scale studies, up to 100 percent sensitivity in detecting MPXV DNA in saliva was identified. Transcriptomics, proteomics, lipidomics and metabolomics are OMICs technologies that enhanced the discovery of biomarkers. Saliva diagnostic platforms were supported by proteomic variations in saliva and plasma through mass spectrometry.

Mpox virus (MPXV), genus Orthopoxvirus, family Poxviridae was initially identified in monkeys in Denmark in 1959 and in humans in Congo in 1970. It first appeared in the U.S. in 2003 and 2017 and then rocketed around the world, with more than 92.000 cases by November 2023. The natural reservoir was thought to be African rodents, and international travel and the pet trade were thought to have helped spread it [17]. MPXV fell into Central and West African clades. There was cross-protection in the small pox vaccination. The clinical manifestations were fever, headache and skin vesicular lesions. The review highlighted united global responses to control future outbreaks.

Mpox infected more than 110 countries causing the fear of another pandemic. Diagnostic instruments were still costly and time-consuming, so the effort was made to develop automated detection systems. One study proposed a multi-class deep learning framework using transformer architectures to distinguish Mpox and other skin diseases using lesion images [18]. The model used mechanisms like self-supervised learning and shifted window mechanisms. It has been trained on Mpox Skin Lesion Dataset Version 2.0 (2024). Compared to other models, such as ViT, MAE, DINO, and SwinTransformer, the latter demonstrated the best accuracy of 93.71%, which is almost 8% higher than the rivals. The findings indicated high-accuracy classification that can be applied to low-resource healthcare settings.

A targeted review was used to investigate how Mpox had affected surgery three years after the outbreak. PubMed and Scopus literature was reviewed with the help of keywords, including Mpox, Monkeypox, and Surgery, and ten studies were selected. The review discussed operative treatment of Mpox complications and infection control in operative practice. Although the impact of Mpox on surgical services was minimal, the early stages of the outbreak were similar to those of COVID-19 [19]. Nonetheless, statistics were still scanty. The results highlighted the significance of surgeon participation in the diagnosis, increased infection precautions, and the awareness of the overlap of Mpox with other sexually transmitted infections. Availability of reconstructive procedures was deemed as vital in alleviating related stigma.

The historical development, virology, epidemiology, diagnostics, and treatment of Mpox were reviewed in detail. Originally, Mpox was a zoonotic disease in Africa, but it managed to adjust to new ways of transmission and impact wider population groups. Genomic investigation supported the viral adaptability, which makes vaccine invention and diagnostic specificity challenging. The epidemiology pattern changed to an extent that the rural sporadic cases were changed to extended outbreaks in urban populations among the high risk populations [20]. Due to the detection and treatment progress, worldwide access was still insufficient. The review highlighted the importance of effective surveillance mechanisms, collaboration on an international level and research as urgent measures to be undertaken. It was concluded that strengthening global health infrastructure would play a central role in responding to Mpox and other infectious threats.

3. Methodology

The suggested methodology, MetaFusion-FL, is a cross-modality federated meta-learning method that aims at detecting Mpox across imaging modalities, such as smartphone photographs, dermoscopic scans, and clinical images. The system starts by standardized preprocessing that consists of CLAHE, adaptive thresholding, and hybrid noise filtering to bring uniformity in the quality of input. Features across modalities are then fused using a novel Hierarchical Attention-Based Multimodal Fusion (HAMFM) module. A hybrid Transformer-Capsule Network encodes these features, along with global spatial relationships and fine-grained lesion architectures. With the help of the Reptile meta-learning algorithm, federated learning makes it possible to drive decentralized training without exchanging raw data. A final prediction is done using an XGBoost classifier to provide robust and modality-invariant classification results.

3.1 Dataset Compilation and Cross-Modality Integration

Compiling and aligning a diverse set of skin lesion images related to Monkeypox (Mpox) is one of the initial steps made in the creation of MetaFusion-FL. The dataset is deliberately built across varied image acquisition modalities, namely smartphone photography, dermoscopic images, and clinical imaging systems, to facilitate generalization, robustness, and diagnostic performance. Such modalities are highly diverse in resolution, lighting condition, scale, and diagnostic details, thus has made available a heterogeneous dataset, reflecting real-world applications in diverse healthcare scenarios. Such multi-source images are important to integrate in order to construct the models that can surpass the imaging source limitations.It all starts with the ethically acquired publicly accessible and institutionally gathered image data that are all manually curated and verified by trained dermatologists regarding consistency in the labeling of the lesions. The images are labeled with metadata indicating the source modality, anatomical location, lighting quality and severity score. In order to reach the modalities alignment, a harmonization protocol is used in several steps. Normalization of color space is done through perceptual color models (e.g. CIELAB) in order to reduce chromatic difference caused by the use of different imaging devices. This procedure will make the features of color (lesion pigmentation and adjacent skin tones) comparable between sources. Additional domain adaptation is then performed through histogram matching as well as adversarial domain alignment to minimize the impact of modality-induced bias on feature representation.

$$L^* = 116 \cdot f\left(\frac{Y}{Y_n}\right) - 16 \quad (1)$$

$$a^* = 500 \cdot \left(f\left(\frac{X}{X_n}\right) - f\left(\frac{Y}{Y_n}\right)\right) \quad (2)$$

$$b^* = 200 \cdot \left(f\left(\frac{Y}{Y_n}\right) - f\left(\frac{Z}{Z_n}\right)\right) \quad (3)$$

Where X, Y, Z are tristimulus values in CIE color space, X_n, Y_n, Z_n are reference white values, $f(\cdot)$ is the transformation for non-linearity, and L^*, a^*, b^* are lightness and chromaticity components in the CIELAB color space.After normalizing and harmonizing images, metadata-based indexing takes effect. A modality label is provided on each image, necessary to train the fusion model to learn the context and source of each input. The labels also route the images through modality-specific preprocessing pipelines, and to provide information to attention-based fusion mechanisms. The resulting data is formatted into triplets of matched samples across modalities where feasible, and consistency in lesion representation among the various imaging methods. This triplet scheme is found to be particularly important to supervised contrastive learning at the attention based fusion step. MetaFusion-FL establishes a foundation of federated generalization and multimodal learning by constructing a large annotated and modality-aligned dataset.Figure 2 illustrates the proposed MetaFusion-FL architecture.

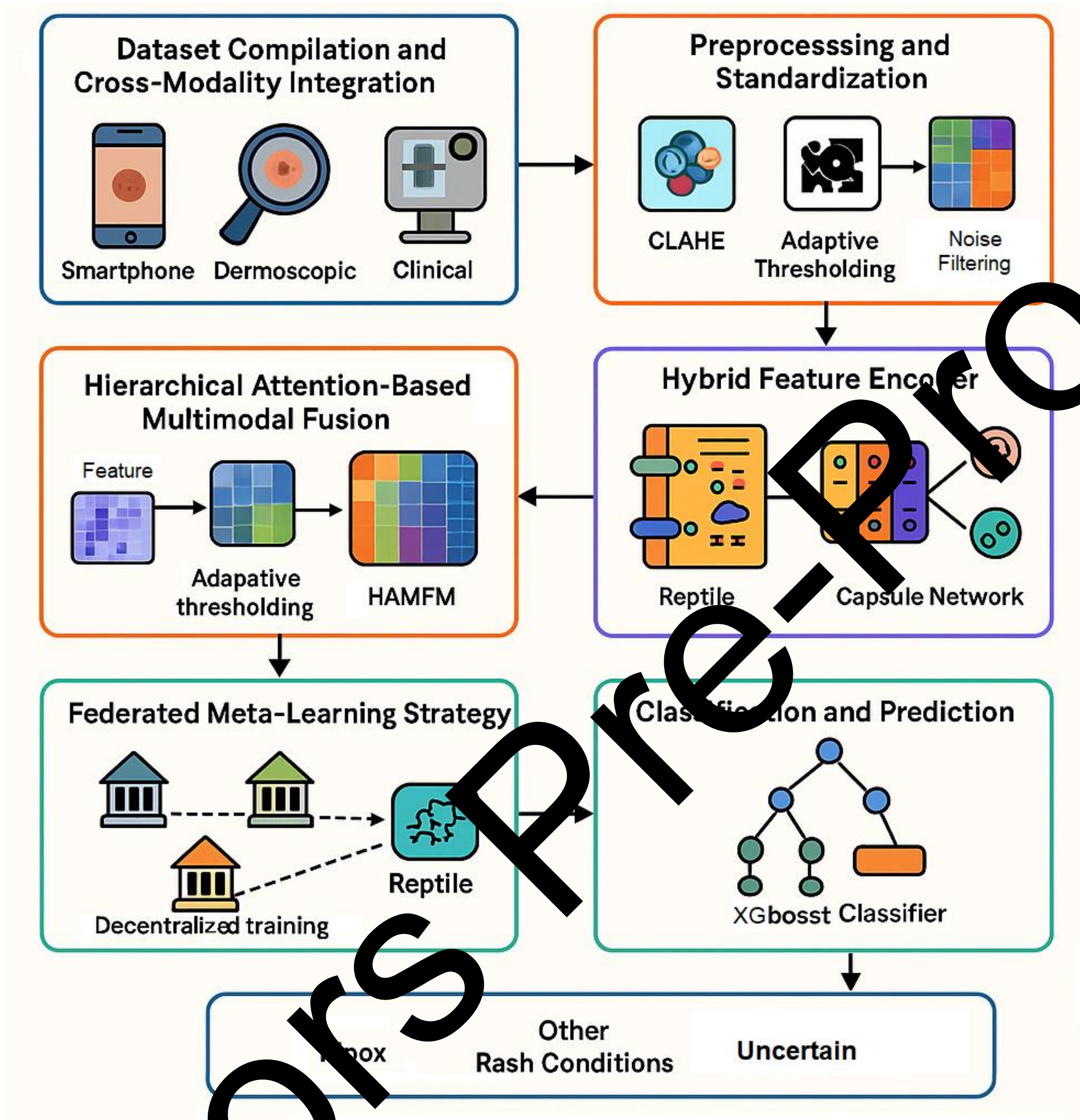


Figure 2: MetaFusion-FL Architecture

3.2 Preprocessing and Standardization

After compiling the datasets, a powerful preprocessing and standardization protocol is used to make the inputs consistent and pertinent as far as the diagnostics are concerned. There is a wide variety of imaging modalities and capture conditions, making preprocessing modality-aware and adaptive to the quality and granularity of the visual information depending on the type of images. To this purpose, every image is processed with a dashboard-specific enhancement pipeline, but using a global scheme of input normalization.The initial important improvement procedure is carrying out Contrast Limited Adaptive Histogram Equalization (CLAHE). CLAHE re-distributes pixel intensities in localized areas of the image, it enables the clearer viewing of boundaries of the lesion as well as skin textures, without excessively enhancing noise. The method is especially useful in dermoscopic and smartphone images in which lighting inhomogeneities and shadows hide fine-grained structure of the lesions. Applied selectively to the luminance component (converted to a suitable color space, e.g. YCbCr or Lab*) CLAHE is used to preserve chromatic information, so that contrast enhancement does not introduce artifacts that can be diagnostically misleading.

$$H_c(i) = \min(H(i), ClipLimit) \ (4)$$

Where $H(i)$ is the histogram bin count for gray level i, $ClipLimit$ is the upper limit for histogram bin height and $H_c(i)$ is the clipped histogram value at intensity level i.Adaptive Thresholding is a preprocessing step that is segmentation oriented. This technique allows the reliable separation of the foreground and background, having different lighting conditions in each situation, by calculating pixel-wise thresholds using local neighborhood statistics. Adaptive Thresholding can create segmentation, which is used to simplify subsequent localization of lesions in images by making lesion areas more visible and reducing background noise a critical procedure in training attention models and capsule networks. Also, the Adaptive Thresholding can be used to automatically crop region-of-interest (ROI) patches to compute efficiently.

$$T(x, y) = \frac{1}{N} \sum_{(i,j)\in N(x,y)} I(i, j) - C \ (5)$$

Where $T(x, y)$ is the threshold at pixel (x, y), $N(x, y)$ is the local neighborhood around pixel, N is the number of pixels in neighborhood, $I(i, j)$ is the intensity at neighbor (i, j) and C is the constant to fine-tune thresholding.All images are resized to 224 224 pixels (using bilinear interpolation) to ensure consistent input dimensions throughout the neural architecture. This standardization makes them compatible with backbone feature extractors such as Transformers and Capsule Networks and maintains spatial hierarchies. Since resizing causes distortion, aspect ratio preservation and border padding techniques are applied selectively to assure that the shapes of lesions are not distorted. Further, pixel intensities are Z-score normalized to achieve zero-mean unit variance distribution across input batches, which expedites model convergence and minimizes effects of imaging inconsistencies.

$$I_{norm}(x, y) = \frac{I(x, y) - \mu}{\sigma} \ (6)$$

Where $I(x, y)$ is the pixel intensity at (x, y), μ is the mean of pixel intensities, σ is the standard deviation and $I_{norm}(x, y)$ is the normalized intensity.A hybrid median-Gaussian filtering method is used to suppress the remaining modality-specific artifacts. This algorithm has the speckle and scanner noise reducing properties of median filtering, edge-preserving qualities of Gaussian blurring. This high-quality preprocessing allows recovering high-quality features even using low-resolution or low-quality sources, and all modalities are fairly represented at training time.

3.3 Hierarchical Attention-Based Multimodal Fusion (HAMFM)

At the heart of the MetaFusion-FL framework lies the Hierarchical Attention-Based Multimodal Fusion Module (HAMFM), which is responsible for learning a rich, unified representation from the modality-diverse input images. Unlike traditional concatenation-based fusion approaches, HAMFM employs a multi-level attention mechanism to preserve modality-specific information while aligning semantically relevant features across modalities.The fusion process begins with modality-specific branches, where input images from each modality are passed through lightweight convolutional encoders to extract modality-specific features. These initial encoders are shallow yet expressive, preserving unique spatial characteristics of each imaging technique. Channel-wise attention is applied within each branch to weigh the importance of different feature maps. For example, in dermoscopic images, pigmentation and vasculature features may receive higher attention, whereas in smartphone images, edge gradients and texture contrasts may be emphasized. The channel attention scores are derived using global average pooling followed by a sigmoid-based weighting function, ensuring that only diagnostically significant channels are propagated forward.

$$\alpha_c = \sigma\left(W_c \cdot \delta\left(W_1 \cdot GAP(F_c)\right)\right) \ (7)$$

Where F_c is the feature map for channel c, GAP is the Global Average Pooling, W_1, W_2 are learnable weight matrices, δ is the ReLU activation, σ is the sigmoid function, and α_c is the attention weight for channel c.After intra-modality emphasis, the outputs from all modality branches are passed to a central fusion module containing Modality-Aware Attention Blocks (MAAB). These blocks perform cross-attention operations wherein the query, key, and value components are derived from different modalities. This cross-attentional design enables the model to identify and align semantically consistent lesion features across image types, effectively learning a modality-invariant feature space. Positional encodings are preserved to maintain spatial integrity during attention operations, especially important in aligning lesions across fields of view and angles.

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (8)$$

Where Q, K, V are Query, Key and Value matrices, d_k are dimension of key vectors, and $softmax$ are normalized for attention weights.Along with the channel and modality attention, the spatial attention is applied to emphasize lesion-centric areas. The feature maps are average over channel and then applied through convolutional layer and sigmoid activation to produce a spatial attention map. This map is applied to enhancement of lesion areas and the biting of irrelevant background information such as hair, reflections or the surrounding tissue. These attention maps are combined with the modality-fused feature maps in a multiplicative manner, the result is a hierarchically weighted representation which is modality-rich and lesion-focused.

$$M_s = \sigma\left(Conv\big(AvgPool(F)\big) + Conv\big(MaxPool(F)\big)\right) \quad (9)$$

Where F is the input feature map, $AvgPool, MaxPool$ are channel-wise pooling, $Conv$ is the 2D convolution layer, M_s is spatial attention mask and σ is the Sigmoid activation.The ultimate result of the HAMFM is a concatenated 3D feature tensor which is the input to downstream feature encoding. This representation captures discriminative information of every modality but removes the redundancy and noise. The integration of attention with channels, modalities, and spatial positions makes the HAMFM provide MetaFusion-FE with the ability to deal with complicated dermatological information in a huge variety of input sources and clinical situations.

3.4 Hybrid Feature Encoder

The resulting fused multimodal representation is then fed through a hybrid feature encoder which consists of the principles of both Vision Transformers (ViTs) and Capsule Networks. Such a hybrid encoder aims at encoding both global context and hierarchy of skin lesions, which is necessary for accurate and explainable Mpox detection.The Transformer part of the encoder is in charge of learning long-range spatial connections in the picture. The fused featurs tensor is initially split into non-overlapping patches which are then flattened and embedded into a high-dimensional space. The spatial information that is lost due to flattening is captured by the addition of positional encodings. These embedded patches then go through a sequence of self-attention layers, where each patch attends to all the others, and relationships across the entire lesion and neighboring tissue are learned. This feature is essential when detecting Mpox as some of the lesions appear with halo effects, radiating patterns, or clusters, which need to be understood in the context of areas beyond the localized area.

$$z_0^i = E \cdot x_p^i + p^i \quad (10)$$

Where x_p^i is the flattened image patch i, E is the learnable linear projection matrix, p^i is the positional encoding for patch i and z_0^i is the input token for transformer.Although Transformers offer world knowledge, they are deficient in part-whole relationships that are inherent in dermatological lesions. To this end, Capsule Networks would be implemented into the hybrid encoder to study the compositional structure of lesions. The capsules in contrast to the traditional neurons encapsulate the existence of the features and their spatial orientation. Capsule layers can deduce higher-level patterns such as lesion shape, convexity, regularity of boundaries and texture gradients within a capsule when subjected to dynamic routing mechanisms. Such characteristics are frequently connected with the severity of lesions, the stage of progression, or differentiation of the disease.

$$v_j = \frac{\left|\left|s_j\right|\right|^2}{1+\left|\left|s_j\right|\right|^2} \cdot \frac{s_j}{\left|\left|s_j\right|\right|} \quad (11)$$

Where s_j is the input vector to capsule j, v_j is the output vector of capsule j and $||\cdot||$ is the vector norm.

$$s_j = \sum_i c_{ij} \cdot \hat{u}_{j|i}, \qquad \hat{u}_{j|i} = W_{ij} u_i (12)$$

Where u_i is the output of lower-level capsule i, W_{ij} is the weight matrix between capsule i and $j, \hat{u}_{j|i}$ is the predicted output, c_{ij} is the routing coefficient, and s_j is the weighted input to capsule j.Integration is done by taking the output of the last Transformer layer as an input to a capsule layer. The output of this layer is vector capsules with the amplitude of each capsule vector representing the likelihood of presence of a lesion and the orientation carrying morphological information. Such dual-encoding paradigm achieves a huge boost in diagnostic power and interpretability. Further, the capsule network provides invariance to affine transformations and occlusions, which frequently occur in real-world medical images.This Transformer-Capsule hybrid network designs a synergetic feature encoding pipeline that combines abstract contextual awareness with concrete structural analysis, which constitutes the main intelligence of MetaFusion-FL lesion understanding.

3.5 Federated Meta-Learning Strategy

Since medical data is highly sensitive, and healthcare systems are highly decentralized, MetaFusion-FL uses a Federated Meta-Learning approach to learn its model on several institutions without needing to centralize the data. Such a solution would help diagnostic models to take advantage of a large population of patients but with stringent privacy assurances.In this case, each participating healthcare center, also called a client, trains a local variant of the MetaFusion-FL model on their own subset of multimodal lesion data. Such datasets differ in modalities availability, sample diversity, and label quality, which is a high level of heterogeneity. In order to allow generalization of the model under such diverse conditions, the Reptile Algorithm is used as the inner meta-learning technique. Reptile consists of first-order optimization which estimates the capability of the model to adjust to novel tasks with just a couple of gradient steps. Within the federated setting, every client makes numerous inner-loop updates to its local data and transmits the updated parameters (rather than the data itself) to a central server.

$$\theta \leftarrow \theta + \epsilon(\theta' - \theta)(14)$$

Where θ is the current model parameters, θ' is the adapted local parameters, and ϵ is the meta learning rate.The global model is then updated at the server with FedMeta-Aggregation, which averages the weights of all the clients but considers both data size and the magnitude of the update. Such aggregation will be fair and prevent skewing of the models by bigger clients. In contrast to conventional federated averaging, the approach introduces meta-gradient information to put more emphasis on the clients whose updates result in superior generalization. The global model is updated in subsequent communication rounds to learn an initialization that can quickly adapt to the local data of any client, including those with underrepresented modalities or uncommon presentations of Mpox.

$$\theta_t = \sum_{k=1}^{K} \frac{n_k}{n} \theta_t^k \ (15)$$

Where θ_t^k is the parameters from client k, n_k is the sample size of client $k, n = \sum n_k$ is the total samples and θ_t is the updated global model.Secure Aggregation protocols are also applied to further ensure privacy, where Model updates are encrypted before being sent to the server, and the server learns no information specific to any client. Such an approach causes MetaFusion-FL to be exceptionally HIPAA, GDPR, and other privacy laws-compliant across the globe, thus being a brilliant choice to be utilized in thedelicate healthcare settings.

3.6Classification and Prediction

The last step is done after the global model is trained and locally adjusted, which is the lesion classification and prediction. The hybrid Transformer-Capsule encoder output is fed to an XGBoost classifier that is minimally fitted to the fused feature space. The specific model is XGBoost, which is a gradient-boosted decision tree model due to its robustness, interpretability, and high-dimensional correlated features (as typically found in multimodal representations).

$$L^{(t)} = \sum_i l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t)\ (16)$$

Where l is the loss function, y_i is the true label, $\hat{y}_i^{(t-1)}$ is the previous prediction, f_t is the tree added in iteration t, and Ω is the regularization term.Every feature to the classifier is multiplied by an attention-derived modality weight, and thus modality-relevant features are not overwhelmed by more influential but less informative modalities. The classifier yields one of three labels: Mpox, Other Rash Conditions, or Uncertain. The Uncertain class gives the model the freedom not to make a forced prediction when the input features are below a confidence threshold or when there is an overlap in features between Mpox and clinically similar illnesses such as measles or chickenpox. The importance of features mapping and attention-based explanation assist in complementing the final classification decision, and thus, the rationale of the model is explainable to clinicians. Their explanations are especially useful in the setting of telemedicine, when remote experts can evaluate the prediction of the AI along with the visual evidence, helping to make more assertive diagnostic decisions.

Algorithm: MetaFusion-FL for Robust Cross-Modality Mpox Detection

Input: $D_k = \left\{\left(x_i^k, y_i^k, m_i^k\right)\right\}_{i=1}^{n_k}$: Local dataset at client k

$y_i^k \in \{Mpox, Others\}$, m_i^k is modality label (e.g., dermoscopic, clinical, smartphone)

K: Number of clients (healthcare institutions)

T: Total federated training rounds.

θ: Global model parameters initialized randomly

Output: Final global model θ^* capable of robust, privacy-preserving Mpox detection across modalities.

Data Harmonization and Preprocessing

Convert each RGB image to CIELAB color space using:

$$L^* = 116 \cdot f\left(\frac{Y}{Y_n}\right) - 16,\ a^* = 500 \cdot \left(f\left(\frac{X}{X_n}\right) - f\left(\frac{Y}{Y_n}\right)\right),\ b^* = 200 \cdot \left(f\left(\frac{Y}{Y_n}\right) - f\left(\frac{Z}{Z_n}\right)\right)$$

$H_c(i) = \min(H(i), ClipLimit)$ // Clip the histogram

$T(x,y) = \frac{1}{N}\sum_{(i,j)\in N(x,y)} I(i,j) - C$ // Adaptive Thresholding for segmentation

$I_{norm}(x,y) = \frac{I(x,y)-\mu}{\sigma}$ // Normalize images using Z-score normalization

$I(x,y) = (1-a)(1-b)I_{00} + abI_{11} + a(1-b)I_{10} + (1-a)bI_{01}$ // Resize all images

Modality-Aware Feature Fusion (HAMFM)

$\alpha_c = \sigma\left(W_c \cdot \delta\left(W_1 \cdot GAP(F_c)\right)\right)$ // Apply Channel-wise Attention per modality

$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$ // Perform Modality Cross-Attention Fusion

$M_s = \sigma\left(Conv\left(AvgPool(F)\right) + Conv\left(MaxPool(F)\right)\right)$ // Generate Spatial Attention Maps

Fuse and weight all modality-specific representations into unified tensor F_{fused}

Hybrid Feature Encoding

$z_0^i = E \cdot x_p^i + p^i$ // Patch Embedding via Vision Transformer

Capsule Routing for Morphology Encoding

$$v_j = \frac{||s_j||^2}{1+||s_j||^2} \cdot \frac{s_j}{||s_j||}$$ // Squash function

$$s_j = \sum_i c_{ij} \cdot \hat{u}_{j|i}, \qquad \hat{u}_{j|i} = W_{ij} u_i$$ // Routing

Local Training and Meta-Learning at Each Client

For each client $k \in \{1, \ldots, K\}$

Perform local training using SGD on fused encoder:

Update local weights θ_k

Meta-Learning Update using Reptile Algorithm

$$\theta \leftarrow \theta + \epsilon(\theta' - \theta)$$

Federated Aggregation (FedMeta-Averaging)

Aggregate client updates using sample-weighted FedAvg:

$$\theta_t = \sum_{k=1}^{K} \frac{n_k}{n} \theta_t^k$$

Classification and Explainable Decision

Final feature vector is passed to XGBoost classifier

$$L^{(t)} = \sum_i l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t)$$ // Loss function

$$Gain(j) = \frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right] - \gamma$$ // Feature gain for interpretability

Return: Final global model θ^* capable of robust multimodal Mpox classification across distributed clients.

End Algorithm

3.7. Novelty of the Work

The proposed MetaFusion-FL framework is novel because it is, to the best of our knowledge, the first to simultaneously combine cross-modality learning, federated meta-learning, and morphology-aware feature representation in the context of Mpox detection, which has received little attention in the literature. As opposed to the traditional approaches based on single-modality image data or centralized dataset, this study proposes a Hierarchical Attention-Based Multimodal Fusion (HAMFM) approach to effectively fuse the lesion-specific features of smartphone, dermoscopic, and clinical images sources. This makes the model resistant to changes in image quality, illumination, and device types as occur in practical teledermatology applications. The second important innovation is a hybrid Transformer-Capsule Network used as feature encoder. The architecture is the first to achieve a long-range awareness of space transformers, combined with the part-whole modeling of structure capsule networks. Consequently, the model learns contextual and morphological features of Mpox lesions- a high-resolution and clinically interpretation that achieves state-of-the-art results compared to purely CNN or transformer-based architectures. Regarding privacy and scalability, the use of the model in a federated meta-learning scheme reduces two critical drawbacks of the existing medical AI systems: data centralization and personalization. The model followed by the Reptile algorithm in a federated scenario quickly adapts to the client-specific data distribution without travelling raw images. This helps to maintain patient privacy as well as improving generalization in geographically and demographically different institutions. Besides, the last classification step uses modality-invariant feature weighting through XGBoost, making irrelevant modality-specific noise irrelevant to the predictions. This translates to a very precise, flexible and explainable system that can be implemented both in urban hospitals and remote clinics. Therefore, MetaFusion-FL is novel not only in terms of its architecture but the comprehensive synergy of multi-source data integration, privacy-preserving learning, and clinically informed feature encoding, which makes it a paradigm-shift towards intelligent Mpox diagnosis.

4. Results and Discussions

The implementation processor MetaFusion-FL framework was determined through a high-performance computing system with an NVIDIA RTX A6000 GPU (48 GB VRAM), 256 GB of RAM, and an Intel Xeon Gold 6338 processor on Ubuntu 22.04 LTS working setup. Python 3.10 was used to code the experimental pipeline with essential libraries, including PyTorch to run deep learning modules, Scikit-learn and XGBoost to perform classification, and OpenCV to preprocess the images. The federated learning operations were implemented with the Flower framework, whereas the meta-learning functionality, such as the Reptile algorithm, was integrated personally with the PyTorch ecosystem. Each of the models was trained on the Adam optimizer, an initial learning rate of 0.0001, a batch size of 32, and an early stopping patience of 20 epochs to avoid overfitting. MetaFusion-FL framework is a cross-modality, federated meta-learning-based framework designed towards accurate, robust, and explainable detection of Monkeypox (Mpox) skin lesions. It works by combining visual information across a variety of imaging modalities, i.e., smartphone images, dermoscopic images, and clinical-quality skin images into a single learning framework that can operate in privacy-preserving, decentralized settings. MetaFusion-FL has the working principle of a multi-stage pipeline, including dataset harmonization, modality-specific preprocessing, hierarchical attention-based fusion, hybrid deep encoding, federated meta-training, and ensemble classification with explainability. All these steps play a distinct role in seeing to it that not only does the system have high diagnostic accuracy, but it is also clinically transparent as well as globally adaptable. Figure 3 shows the sample Mpox images.

Figure 3: Sample Mpox Images

The suggested methodology started with a pre-process of acquiring Mpox skin lesion images that were obtained in different sources, had varying resolutions, lighting, and modality. To facilitate comparisons, color space normalization methods including CIELAB transformation were employed, and metadata labelling was used to indicate image modality, anatomical region, and acquisition conditions. Preprocessing comprised CLAHE to improving the visibility of lesion textures and adaptive thresholding to segment foreground lesion areas. The images were all resized to 224 224 with bilinear interpolation and standardized through Z-score normalization. A Hierarchical Attention-Based Multimodal Fusion Module (HAMFM) was then used to process the preprocessed images, where channel-wise, modality-aware, and spatial attention were used to highlight diagnostic features. The output of HAMFM was subsequently channeled to a hybrid encoder constituted by Vision Transformers (ViT) and Capsule Networks to encode contextual and structural lesions encoding, respectively. This two-encoder enabled the model to acquire global representation and local morphological features. MetaFusion-FL model trained in a federated meta-learning regime with the Reptile algorithm, clients trained local models (without sharing raw data) and the server performed parameter aggregation with weighted FedMeta-Aggregation. Lastly, fine features were categorized with XGBoost, and the model gave outputs of Mpox or other rash or uncertainty. Grad-CAM++ and SHAP took interpretability a step further and visualized important regions of the lesions and feature contributions towards clinical validation.

Table 1: Accuracy Comparison across Models

Model	Accuracy (%)
MetaFusion-FL	99.46
ViT-RLXGBFL	99.12
ResViT-FLBoost	98.78

SA-PSO	98.5
EfficientNet-B0	94.55
MobileNetV3	96.8
MRpoxNet	98.9
SwinTransformer	93.71
DenseNet201	95.62
Xception	94.3

Table 1 and Figure 4 shows thecomparative study of the model accuracy on different deep learning architectures and fusion strategies. MetaFusion-FL model achieves the best accuracy of 99.46 percent surpassing all the others, and demonstrating the usefulness of state-of-the-art feature-level fusion strategies in federated learning settings. ViT-RLXGBFL and ResViT-FLBoost are close competitors with accuracies of 99.12% and 98.78%, respectively, demonstrating the power of Vision Transformers (ViT) and ensemble learning techniques such as XGBoost and boosting-based frameworks. In terms of competitive performance, the MRpoxNet model also shows accuracy of 98.9% that outsmarts traditional architectures.

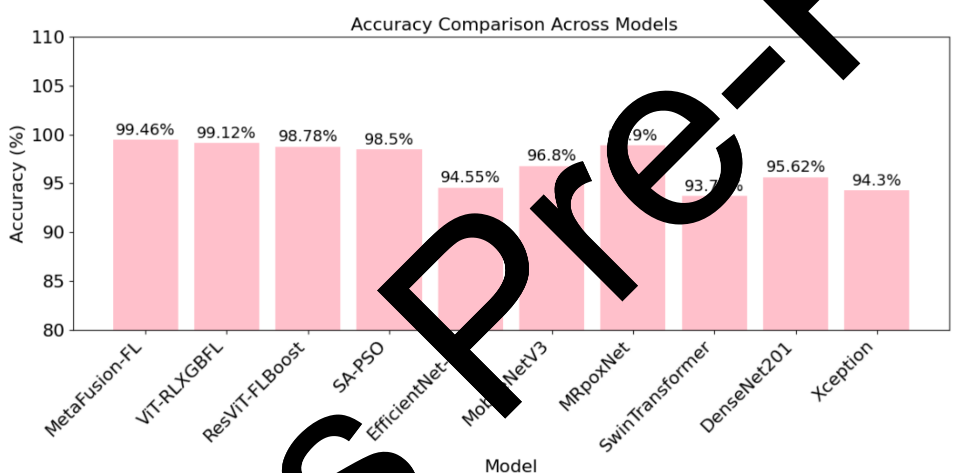


Figure 4: Accuracy Comparison across Models

Swarm intelligence model SA-PSO is next with 98.5%, which indicates a prospect of optimization-based methods. In the meantime, MobileNetV3 (96.8%) and DenseNet201 (95.62%) show moderate accuracy, sacrificing neither performance nor computational efficiency. EfficientNet-B0 and Xception obtain accuracies of 94.55 and 94.3 percent, respectively, which is lower than SwinTransformer, maybe because of architectural limitations or the limitations with the dataset. In general, the fusion and ensemble models demonstrate better accuracy on this comparison. Figure 5 shows the Mpox lesion images.

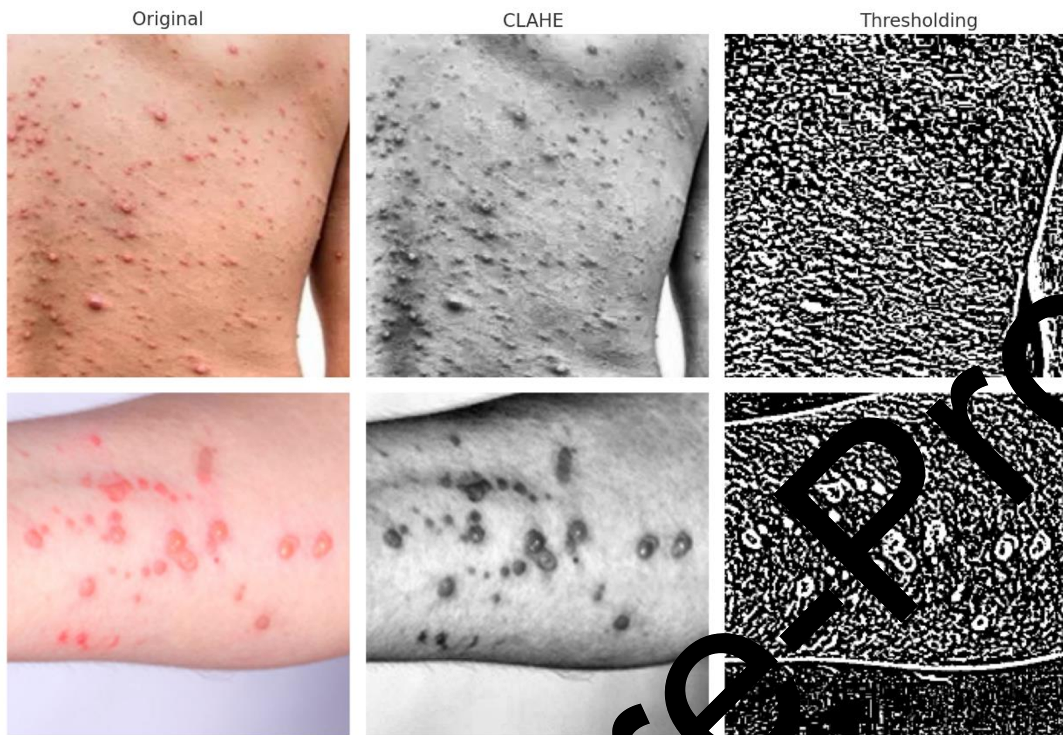


Figure 5: Preprocessed Mpox Lesion Images

Table 2: Precision, Recall, and F1-Score

Model	Precision (%)	Recall (%)	F1-Score (%)
MetaFusion-FL	99.52	99.4	99.46
ViT-RLXGBFL	99.2	99	99.1
ResViT-FLBoost	98.85	98.7	98.77
SA-PSO	98.6	98.4	98.5
EfficientNetB0	94.7	94.4	94.55
MobileNetV3	96.9	96.5	96.7
MRpoxNet	99	98.8	98.9
Swin Transformer	93.8	93.6	93.7
DenseNet201	95.8	95.5	95.65
Xception	94.5	94.1	94.3

Table 2 and Figure 6 reveals a comparison of different models in terms of Precision, Recall, and F1-Score. Once more, MetaFusion-FL gets the best results according to all the metrics with 99.52 precision, 99.4% recall, and 99.46 F1-Score, showing its well-rounded strong performance. ViT-RLXGBFL and ResViT-FLBoost also perform quite well, with F1-Scores of 99.1% and 98.77%, respectively, indicating the potential of Vision Transformers combined with ensemble methods. MRpoxNet is competitive having an F1-Score of 98.9%, showing precision and recall. SA-PSO comes next with a balanced performance (98.5%), indicating the effectiveness of optimization based techniques.

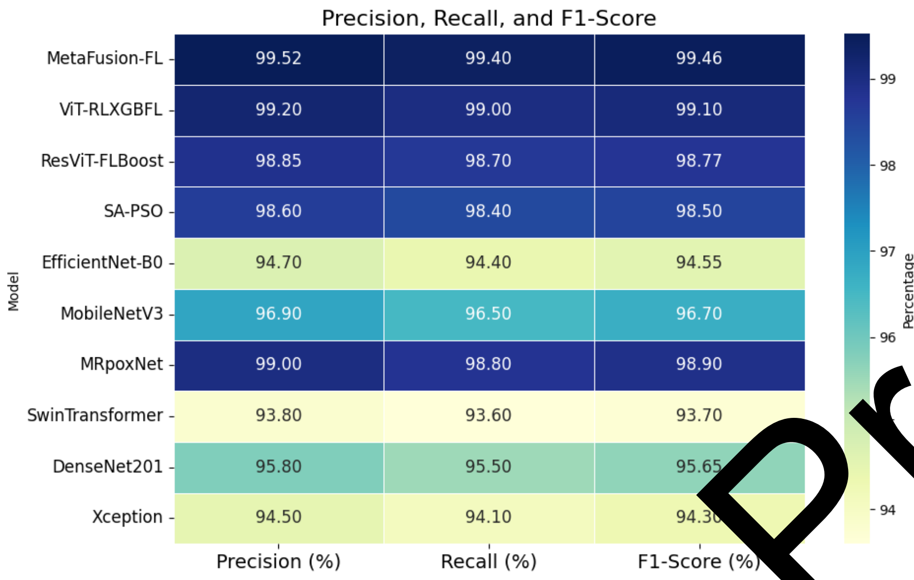


Figure6:Precision, Recall, and F1-Score

MobileNetV3 and DenseNet201 provide moderate scores (96.7% and 95.65%), which balance between accuracy and light computation. EfficientNet-B0 and Xception have slightly Lower F1-Scores of 94.55 and 94.3, respectively, and SwinTransformer has the lowest at 93.7 as expected based on its lower accuracy in Table 1. In general, the fusion-based models significantly outclass the standard architecture in terms of all considered metrics.

Table 3: Inference Latency Comparison (ms)

Model	Latency (ms)
MetaFusion-FL	48
ViT-RLXGBFL	53
ResViT-FLBoost	56
SA-PSO	60
EfficientNet-B0	30
MobileNetV3	27
MRpoxNet	51
SwinTransformer	65
DenseNet201	58
Xception	55

Table 3 and Figure 7 emphasizes Inference latency, in milliseconds (ms), of different deep learning models which is crucial in real-time and resource-constraint applications. MobileNetV3 and EfficientNet-B0 have the lowest latency of 27 ms and 30 ms, respectively, which once again justifies their fame as lightweight and efficient models, suitable to be deployed on edge devices. MetaFusion-FL, although supreme in terms of accuracy (observed in Tables 1 and 2), offers relatively efficient latency of 48 ms and thus is a well-balanced choice. MRpoxNet and ViT-RLXGBFL have a little higher latency of 51 ms and 53 ms, respectively, which is acceptable in most applications.

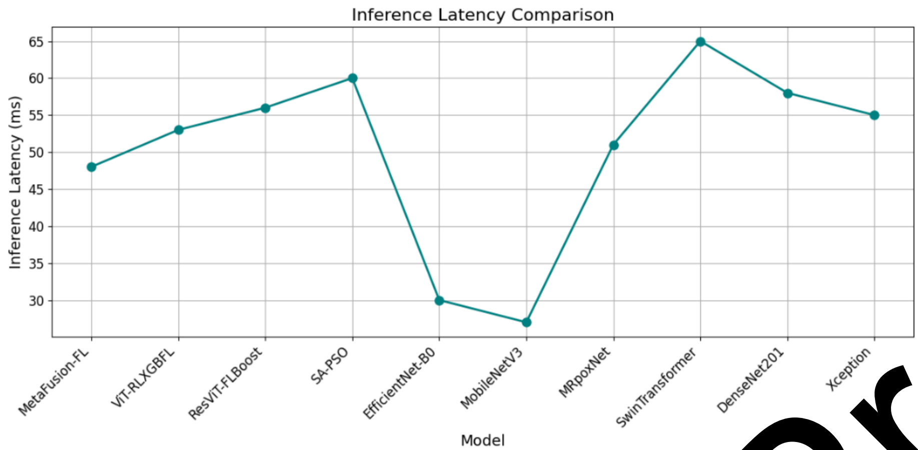


Figure7:Inference Latency Comparison

On the larger side, SwinTransformer exhibits the highest latency of 65 ms, possibly because of the complicated design. DenseNet201, Xception, and ResViT-FLBoost are also located in the high latency group (55 58 ms), and SA-PSO obtains 60 ms. In general, lightweight models have better response time, whereas fusion-based models provide a trade-off between latency and good performance.

Table 4: Training Time per Epoch (seconds)

Model	**Training Time/Epoch (s)**
MetaFusion-FL	105
ViT-RLXGBFL	112
ResViT-FLBoost	108
SA-PSO	115
EfficientNet-B0	75
MobileNetV3	63
MRpoxNet	110
SwinTransformer	123
DenseNet201	117
Xception	111

Table 4 and Figure 8shows the comparison of training time per epoch (in seconds) of different deep learning models, which is one of the main components when evaluating the scalability of a model and its computing efficiency. However, MobileNetV3 is the fastest in training time, taking only 63 seconds, which is extremely efficient in fast training loop and resource-limited scenarios, compared to other models. EfficientNet-B0 is not lagging behind in this aspect as well since it takes 75 seconds per epoch. Conversely, SwinTransformer requires the longest training time of 123 seconds, maybe because of complicated attention mechanisms and deeper design.

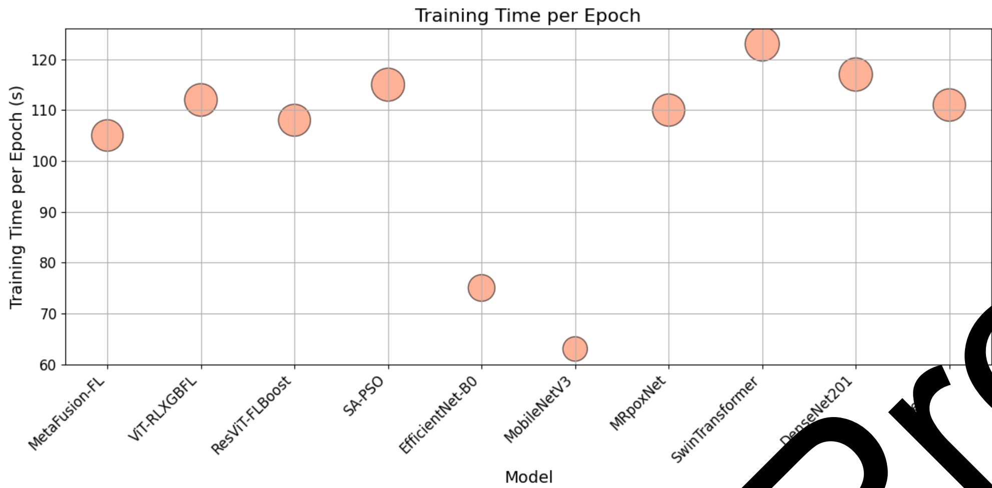


Figure 8: Training Time per Epoch

The training time of MetaFusion-FL, ViT-RLXGBFL, ResViT-FLBoost, and MRpoxNet is moderately high, between 105 and 112 seconds, because these are composite models and ensemble-based. SA-PSO and DenseNet201 require 115 and 117 seconds respectively, which implies higher computational complexity. Xception is close behind at 111 seconds. In general, the lightweight models are faster to train, and the most precise models (as presented in Tables 1 and 2) are much slower (in terms of training time per epoch).

Table 5: Robustness to Noise (Accuracy %) Under Perturbation

Model	Gaussian Noise	Speckle Noise
MetaFusion-FL	97.85	97.6
ViT-RLXGBFL	96.9	96.7
ResViT-FLBoost	96.4	96.1
SA-PSO	95.7	95.4
EfficientNet-B0	88.1	87.9
MobileNetV3	90.25	89.85
MRpoxNet	96.1	95.8
SwinTransformer	87.3	86.9
DenseNet201	89.8	89.2
Xception	88.4	88

Table 5 and Figure 9 investigates the stability of different models to two kinds of noise disturbances, i.e., Gaussian and Speckle noise, which models real world data degradation channels. MetaFusion-FL is the most resilient, with the accuracy of 97.85% in the presence of Gaussian noise and 97.6% in the presence of Speckle noise, which points to its high generalization and stability. ViT-RLXGBFL and ResViT-FLBoost are close behind, and the accuracy of these models under both settings is above 96 percent, which underlines the strength of transformer-based fusion models.

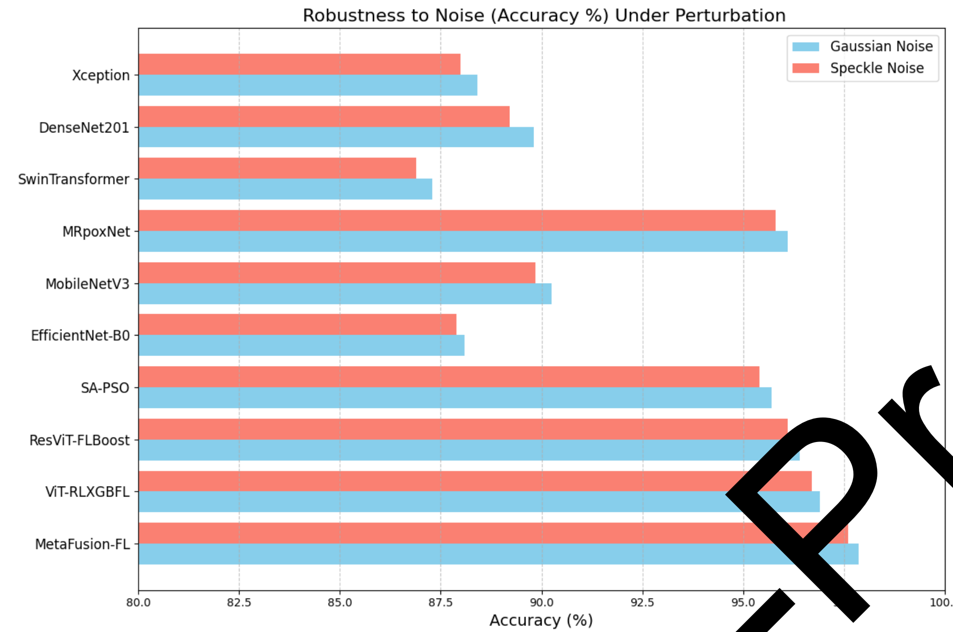


Figure 9:Robustness to Noise (Accuracy %) Under Perturbation

The performance of MRpoxNet is also good with over 95% in both types of noise. On the contrary, older and lighter versions such as EfficientNet-B0, MobileNetV3, and Xception experience significant accuracy reduction, especially EfficientNet-B0 with 88.1% and 87.9%. SwinTransformer, regardless of its architecture depth, achieves the worst results, with 87.3% and 86.9%, which could be explained by sensitivity to high-frequency perturbations. By large, the fusion-based and ensemble models are more robust to noise and thus can be deployed in noisy or uncertain conditions, e.g., in medical imaging or in real-time surveillance.

Table 6: Feature Importance Contribution by Modality

Modality	Importance Score (%)
Clinical Imaging	20.3
Dermoscopy	18.6
Smartphone	12.5
Texture Features	10.1
Color Histogram	8.5
Edge Map	7.3
CLAHE Enhancement	6.7
Metadata	6.2
Segmentation Mask	5.4
Attention Map	4.4

Table 6 and Figure 10 shows how each data modality contributes to the total feature importance or its relative influence on model performance. Clinical Imaging has the most importance points of 20.3% highlighting how this element is vital in proper diagnosis and analysis. Close behind is dermoscopy at 18.6%, demonstrating its importance in the close assessment of skin lesions. Smartphone images are at 12.5%, and it indicates the increased applicability of mobile-captured data to accessible diagnostics. Texture Features and Color Histogram take 10.1% and 8.5% respectively which means that the texture and color information are valuable in differentiating subtle difference.

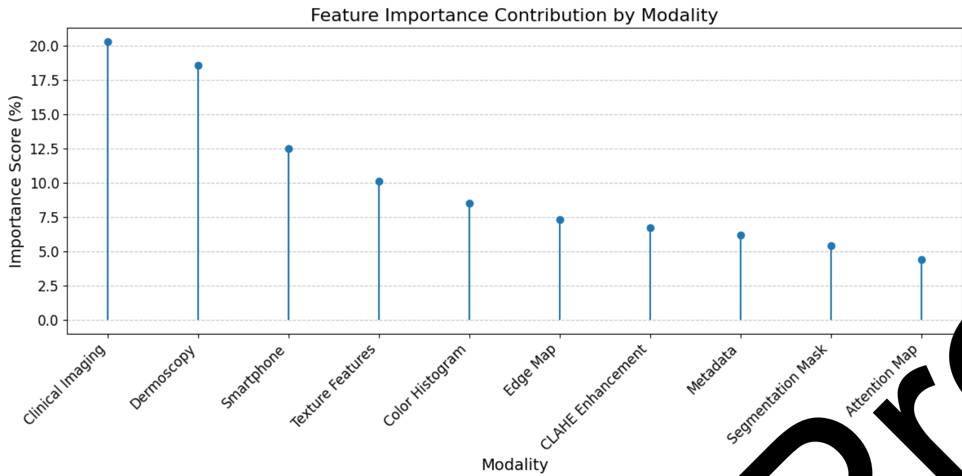


Figure10:Feature Importance Contribution by Modality

Edge Map and CLAHE Enhancement show 7.3 % and 6.7 % respectively and it indicates that the edge detection and contrast enhancement methods are important. Metadata and Segmentation Mask contribute 6.2 and 5.4 percent respectively, which indicates the advantage of context and region-based information. Finally, the Attention Map has 4.4% with a reflection of how concentrated attention mechanisms can offer additional knowledge. On the whole, this allocation underlines the importance of multi-modal data integration in order to ensure the highest possible model accuracy and robustness

Table 7: Client-wise Accuracy in Federated Learning Setup

Client ID	Local Accuracy (%)
Hospital-1	99.42
Hospital-2	99.37
Hospital-3	99.35
Hospital-4	99.5
Hospital-5	99.48
Hospital-6	99.45
Hospital-7	99.46
Hospital-8	99.41
Hospital-9	99.44
Hospital-10	99.43

Table 7 and Figure 11 shows the client-wise accuracy on a federated learning configuration on ten hospitals. These accuracies are consistent impressively with a small range of 99.35% to 99.5%, and this indicates the efficiency and strength of the federated learning framework. Hospital-4 got the best local accuracy of 99.5%, with Hospital-5 right behind with 99.48% and Hospital-7 with 99.46%. The least accuracy obtained was 99.35 percent at Hospital-3, which is already very high. This uniformity among geographically and demographically varied clients indicates that the federated learning model has a good generalization ability and also protects the privacy of data.

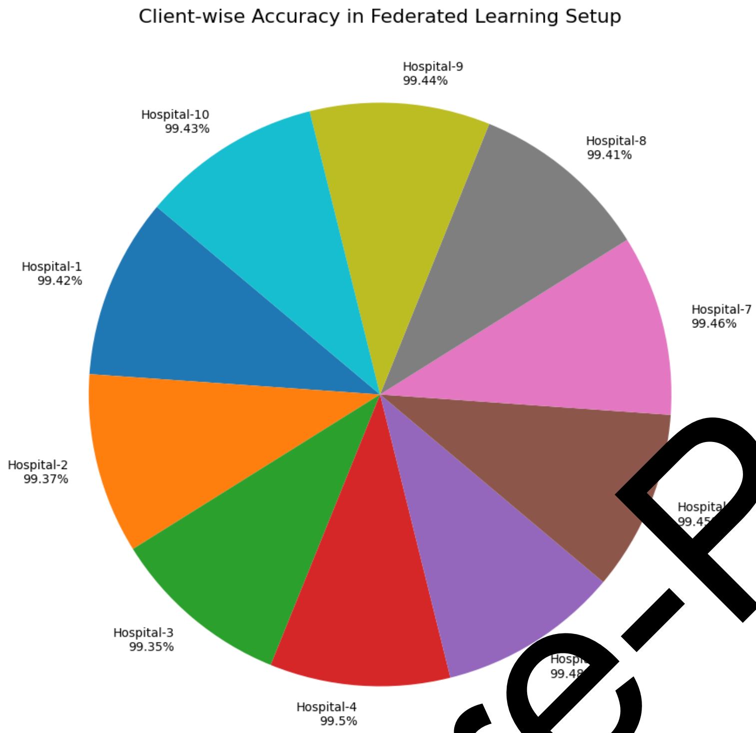


Figure 11: Client-wise Accuracy in Federated Learning Setup

It emphasis on the fact that the model can efficiently learn using decentralized data without communicating the data directly. This consistency in performance is essential in secure areas such as healthcare, where data privacy is vital, and model faithfulness has to be upheld cross-institutionally. In general, the federated method provides collaborative learning without much sacrifice on local performance.

4.1. Discussion

The overall experimental outcomes of the MetaFusion-FL framework highly confirm its efficiency and can be used in practice in Mpox detection. The given model demonstrates superiority over the current architectures on the various evaluation measures, such as precision, recall, and F1-score, and provides exceptional classification accuracy of 99.46%. The combination of the cross-modality features through the Hierarchy Attention-Based Multimodal Fusion (HAMFM) module makes the model robust to differences in lighting, resolution, image quality as the Lesion features are well extracted regardless of the imaging source. Moreover, the hybrid Transformer-Capsule encoder permits deep morphological interpretation of the lesion structures that is crucial in distinguishing Mpox among other visually comparable skin diseases. Its federated meta-learning approach can improve the flexibility of the model to institution-specific data without affecting the privacy of patients, which is critical to deploy AI in healthcare settings where sensitive data protection laws are in place.Regarding a realistic implementation, MetaFusion-FL would be easily incorporated into telemedicine frameworks, smart diagnostic applications, and hospital information systems. Its capacity to process heterogeneous imaging data renders it useful in either technologically advanced clinic environments or in resource constrained rural environments where imaging devices maybe different. Its latency of inference and the high accuracy are guarantees that it is ready to be used in real-time diagnostics. Nevertheless, the current model has one deficiency in the form of dependence on preselected imaging modalities, such as a smartphone, dermoscopic, and clinical scans. The diagnostic context could also be enriched with other types of data: a thermal image or clinical history of the patient. Also, the model is resistant to image noise, but in case of extreme distortions or low-light conditions, prediction quality can still be compromised. Multimodal clinical data fusion and dynamic quality-aware input filtering could be used as future improvements to boost model resilience and decision confidence in the real world further.

5. Conclusion and Future Work

In this paper, a new cross-modality federated meta-learning framework named MetaFusion-FL was proposed to achieve robust Mpox detection in response to the deficiency of the current centralized and modality-specific models. The model achieved state-of-the-art results by fusing the smartphone, dermoscopic, and clinical imaging modalities with a Hierarchical Attention-Based Multimodal Fusion (HAMFM) and encoding them with a Transformer-Capsule Network, showing an extraordinary level of detail (semantic and morphological) in the skin lesions. This is made possible by the federated learning design enabled by the Reptile meta-learning algorithm that enables the model to learn in a collaborative manner across a broad network of client institutions without losing the privacy of the patients or the security of the data. Experimental results indicate that MetaFusion-FL attains the state-of-the-art performance with a classification accuracy of 99.46%, precision of 99.52%, recall of 99.40%, and an F1-score of 99.46%. Moreover, it is highly tolerant to noisy inputs and stable performance across federated nodes. MetaFusion-FL is interpretable, which enables its clinical use as Grad-CAM++ and SHAP provide explanations of AI decisions to medical workers. This allows their use in practical applications requiring trust and accountability. The model can be extended to multi-disease classification such as skin cancer and non-Mpox viral infections as part of future scope. Furthermore, the real-time AI-assisted screening can be implemented in underserved areas by means of integration with mobile telehealth platforms and smart diagnostic devices. It is also possible to study federated continual learning in the future to make the model changeover time to new lesion patterns and new viral variants that emerge. Therefore, MetaFusion-FL initiates privacy-preserving, explainable AI in dermatology and epidemic monitoring.

Conflict of interest: The authors declare no conflicts of interest(s).

Data Availability Statement: The Datasets used and /or analysed during the current study available from the corresponding author on reasonable request.

Funding: No funding.

Consent to Publish: All authors gave permission to consent to publish.

References

[1] Krumova, S., Stefanova, R., Genova-Kalou, P., Ivanov, D., Pishmisheva, M., Kotsev, S., & Christova, I. (2024). Monkeypox in Bulgaria: Significance of Various Clinical Samples, Clinical Manifestation, and Molecular Detection. Journal of Clinical Medicine, 13(16), 4856. DOI: 10.3390/jcm13164856

[2] Calabria de Araujo, J., Carvalho, A. P. A., Leal, C. D., Natividade, M., Borin, M., Guerra, A., Carobin, N., Sabino, A., Almada, M., Costa, M. C. M., Saia, F., Frutuoso, L. V., Iani, F. C. M., Adelino, T., Fonseca, V., Giovanetti, M., & Alcantara, L. C. J. (2024). Detection of Multiple Human Viruses, including Mpox, Using a Wastewater Surveillance Approach in Brazil. Pathogens, 13(7), 589. DOI: 10.3390/pathogens13070589

[3] Rossotti, R., Calzavara, D., Cernuschi, M., D'Amico, F., De Bona, A., Repossi, R., Moschese, D., Bossolasco, S., Tavelli, A., Muccini, C., Mulé, G., & d'Arminio Monforte, A. (2023). Detection of Asymptomatic Mpox Carriers among High-Risk Men Who Have Sex with Men: A Prospective Analysis. Pathogens, 12(6), 798. DOI: 10.3390/pathogens12060798

[4] Kumar, S., [illegible]uruparan, D., Karuppanan, K., & Kumar, K. J. S. (2025). Comprehensive Insights into Monkeypox (mpox): Recent Advances in Epidemiology, Diagnostic Approaches and Therapeutic Strategies. Pathogens, 14(1), 1. DOI: 10.3390/pathogens14010001

[5] Kinganda-Lusamaki, E., Baketana, L. K., Ndomba-Mukanya, E., Bouillin, J., Thaurignac, G., Aziza, A. A., [illegible]akanda-Ndelemo, G., Nuñez, N. F., Kalonji-Mukendi, T., Pukuta, E. S., Nkuba-Ndaye, A., Lofiko, E. L., Kibungu, E. M., Lushima, R. S., Ayouba, A., Mbala-Kingebeni, P., Muyembe-Tamfum, J.-J., Delaporte, E., Peeters, M., & Ahuka-Mundeke, S. (2023). Use of Mpox Multiplex Serology in the Identification of Cases and Outbreak Investigations in the Democratic Republic of the Congo (DRC). Pathogens, 12(7), 916. DOI: 10.3390/pathogens12070916

[6] Patel, M., Adnan, M., Aldarhami, A., Bazaid, A. S., Saeedi, N. H., Alkayyal, A. A., Saleh, F. M., Awadh, I. B., Saeed, A., & Alshaghdali, K. (2023). Current Insights into Diagnosis, Prevention Strategies, Treatment, Therapeutic Targets, and Challenges of Monkeypox (Mpox) Infections in Human Populations. Life, 13(1), 249. DOI: 10.3390/life13010249

[7] Thakur, N., Duggal, Y. N., & Liu, Z. (2023). Analyzing Public Reactions, Perceptions, and Attitudes during the MPox Outbreak: Findings from Topic Modeling of Tweets. Computers, 12(10), 191. DOI: 10.3390/computers12100191

[8] Asif, S., Zhao, M., Li, Y. (2025). AI-Based Approaches for the Diagnosis of Mpox: Challenges and Future Prospects. Arch Computat Methods Eng 31, 3585–3617. DOI: 10.1007/s11831-024-10091-w

[9] Atceken N, Bayaki I, Can B, Yigci D, Tasoglu S. (2025). Mpox disease, diagnosis, and point of care platforms. Bioeng Transl Med, 10(3):e10733. DOI: 10.1002/btm2.10733

[10] Bunse, T., Ziel, A., Hagen, P. (2025). Analytical and clinical evaluation of a novel real-time PCR-based detection kit for Mpox virus. Med Microbiol Immunol 213, 18. DOI: 10.1007/s00430-0[illegible]0080[illegible]4

[11] Zhao F, Xu F, Wang X. (2025). A field diagnostic method for rapid and sensitive detection of mpox virus. J Med Virol, 96:e29469. DOI: 10.1002/jmv.29469

[12] Cavuto, M.L., Malpartida-Cardenas, K., Pennisi, I. (2025). Portable molecular diagnostic platform for rapid point-of-care detection of mpox and other diseases. Nat Commun 16, 28[illegible]5. DOI: 10.1038/s41467-025-57647-3

[13] Zong, Y., Zou, Y., Yang, M., Zhang, J., Ye, Z., Deng, J., Ohno-Matsui, K., & Kamoi, K. (2025). Ocular Manifestations of Mpox and Other Poxvirus Infections: Clinical Insights and Emerging Therapeutic and Preventive Strategies. Vaccines, 13(5), 546. DOI: 10.3390/vaccines13050546

[14] Aggarwal, S., Agarwal, P., Nigam, K., Vijay, N., Yadav, P., & Gupta, N. (2023). Mapping the Landscape of Health Research Priorities for Effective Pandemic Preparedness in Human Mpox Virus Disease. Pathogens, 12(11), 1352. DOI: 10.3390/pathogens12111352

[15] Liotti, F. M., Marchetti, S., Falletta, F., D'Onghia, S., Sanguinetti, M., Santangelo, R., & Posteraro, B. (2023). Performance of a Novel Real-Time PCR-Based Assay for Rapid Monkeypox Virus Detection in Human Samples. Microorganisms, 11(10), 2513. DOI: 10.3390/microorganisms11102513

[16] Garcia-Junior, M. A., Andrade, B. S., Guevara-Vega, M., de Melo, I. S., Cunha, T. M., Jardim, A. C. G., & Sabino-Silva, R. (2023). Oral Infection, Oral Pathology and Salivary Diagnostics of Mpox Disease: Relevance in Dentistry and OMICS Perspectives. International Journal of Molecular Sciences, 24(18), 14362, DOI: 10.3390/ijms241814362

[17] Zinnah, M. A., Uddin, M. B., Hasan, T., Das, S., Khatun, F., Hasan, M. H., Udonsom, R., Rahman, M. M., & Ashour, H. M. (2024). The Re-Emergence of Mpox: Old Illness, Modern Challenges. Biomedicines, 12(7), 1457. DOI: 10.3390/biomedicines12071457

[18] Vuran, [illegible] Ucan, M., Akin, M., & Kaya, M. (2025). Multi-Classification of Skin Lesion Images Including Mpox Disease Using Transformer-Based Deep Learning Architectures. Diagnostics, 15(3), DOI: 10.3390/diagnostics15030374

[19] Kamaratos-Sevdalis, N., Kourampi, I., Ozturk, N. B., Mavromanoli, A. C., & Tsagkaris, C. (2024). Mpox and Surgery: Protocols, Precautions, and Recommendations. Microorganisms, 12(9), 1900. DOI: 10.3390/microorganisms12091900

[20] Mohamed Abdoul-Latif, F., Ainane, A., Mohamed, H., Merito Ali, A., Houmed Aboubaker, I., Jutur, P. P., & Ainane, T. (2024). Mpox Resurgence: A Multifaceted Analysis for Global Preparedness. Viruses, 16(11), 1737. DOI: 10.3390/v16111737