MADRNet: Morphology-Aware Dual-path Reversible Network for Sperm Classification

Fan Yang, Jingzhang Sun, Honglan Huang*, Liang Zhang and Jiheng Zhang

***Abstract*— Sperm morphology analysis plays a crucial role in the clinical diagnosis of male infertility. However, manual evaluation is inherently subjective, and inconsistencies in diagnostic criteria may compromise accuracy. Some existing sperm image classification models are introduced but requiring manual intervention. Most models lack of consideration of alignment between computational classification and WHO sperm morphology standards. To address these challenges, we propose an innovative morphology-aware dual-path reversible network (MADRNet) in designing our model. We integrate key biomarkers, such as head aspect ratio and acrosomal integrity, both of which are crucial for clinical sperm assessment, into the network. Particularly, the network utilizes a dual-path attention mechanism, incorporating both parallel spatial and channel attention, while embedding the acrosome anatomical constraint within the channel attention. To further enhance the alignment of our model with the WHO standards, we develop a dynamic loss function considering head aspect ratio constraint. Further, we employ a reversible architecture to enable the model to preserve more microscopic details while reducing GPU memory consumption. Experiments on the HuSHeM dataset demonstrate that the model achieves an accuracy of 96.3% and an F1 score of 96.8%. Meanwhile, the model maintains a real-time processing speed of 32ms per image, providing a precise and efficient solution for clinical sperm screening. The implementation source code and the underlying dataset are available at https://github.com/fanyangZK/MADRNet.**

***Index Terms*— Sperm Classification, Reversible Architecture, Dual-path Attention Mechanism, Morphology-Aware, WHO Standards**

I. Introduction

MALE factors contribute to 30%-50% of global infertility cases, with sperm morphology assessment being a critical diagnostic indicator in assisted reproductive technologies [1]. The 6th World Health Organization (WHO) edition manual [1] emphasizes that strict criteria for sperm morphology assessment (e.g., acrosome integrity and nuclear vacuole size) are critical, as deviations from these standards may lead to a >90% reduction in fertilization efficacy. However, at present, sperm morphology analysis faces many difficulties.

Clinically, the interpretation of microscopic images by experts has the problems of low efficiency and strong subjectivity [2], [3]. Moreover, large variation and poor consistency in the sperm morphology assessment exist even though experts share the WHO standards, mainly because experts have different experience and perspectives [4]. Deviation from WHO-defined morphological standards will markedly reduce fertilization success rates, thereby increasing the risk of misdiagnosis and suboptimal treatment outcomes.

Deep learning in recent years have significantly advanced the field of automated human sperm image classification. To extract human sperm morphological features, previous studies [5], [6], [8], [20], [27], [53] have attempted to leverage transfer learning of general image classification models. Yet these approaches fail to adequately model the anisotropic texture characteristics inherent in sperm images. Attention mechanisms enhance fine-grained image classification by adaptively focusing on critical features while maintaining global structural relationships, enabling precise differentiation of subtle morphological abnormalities [28], [31], [37], [38]. However, traditional sequential attention mechanisms tend to lose microscopic features and suffer from gradient decay due to excessive path length. To resolve this, we propose a parallel dual-path mechanism that integrates spatial attention and channel attention in parallel, effectively preserving detailed features while mitigating gradient decay.

Little attention has been paid to medical prior knowledge [32], [33] in the field of sperm classification. To enhance WHO compliance, we innovatively introduce acrosome anatomical and head aspect ratio constraints, integrating the former into the channel attention mechanism and embedding the latter in the hybrid loss function. Acrosome integrity affects head contour smoothness and edge continuity, while head width ratio serves as a geometric standard for elliptical contours [35]. Particularly, a deviation of the head aspect ratios outside 1.5 ± 0.3 often indicates abnormal morphologies (e.g., tapered or pyriform) [1]. The introduced constraints guide our model to focus on the acrosome area and penalize sperm with parameters deviating from the standard range, thereby improving the model's ability to distinguish between different categories of sperm.

This work was supported by the Hainan Provincial Natural Science Foundation of China No. 824QN380, Hainan Province Key R&D plan project (No. ZDYF2024GXJS030) and the HK RGC General Research Fund (Nos. 16208120 and 16214121).

Fan Yang, Jingzhang Sun and Liang Zhang are with the School of Cyberspace Security (School of Cryptology), Hainan University, Haikou 570228, China. Liang Zhang is also with Department of Industrial Engineering and Decision Analytics, Hong Kong University of Science and Technology. (e-mail:yfmodify@gmail.com; jingzhangsun@hainanu.edu.cn; zhangliang@hainanu.edu.cn).

Honglan Huang is with the Key Laboratory of Reproductive Health Diseases Research and Translation of Ministry of Education, the First Affiliated Hospital, Hainan Medical University, Haikou 571101, Hainan, China (e-mail: huang-honglan@163.com).

Jiheng Zhang is with the Department of Industrial Engineering and Decision Analytics, Hong Kong University of Science and Technology. (e-mail: jiheng@ust.hk).

*Corresponding author

In addition, the loss of detailed information due to network-level compression is a prevalent issue in deep learning networks [7]. This problem is especially obvious in microscopic image processing, where key details are often lost. To address this challenge, a reversible architecture has been introduced [24], [25], [29]. By combining the parallel dual-path attention mechanism with reversible architecture, the proposed model enables lossless feature reconstruction while preserving microscopic morphological characteristics. Furthermore, the reversible architecture eliminates the need to store intermediate activation values, significantly reducing GPU memory consumption. This allows the model to process higher-resolution images, meeting WHO requirements for microscopic detail resolution. Thus, the proposed model is tailored for sperm classification with parameter scale compression and real-time computational efficiency.

Beyond architectural advancements, loss function optimization represents another critical strategy [8], [41], [42], [44]. Traditional cross-entropy loss with rigid classification boundaries tends to over-penalize borderline samples, failing to reflect continuous morphological variations. To address this, we design a hybrid loss function incorporating triplet loss to enhance discriminative feature mining and improve robustness for borderline samples.

Summarily, we put forward a morphology-aware dual-path reversible network (**MADRNet**) for human sperm classification. The main contributions are:

- We propose a parallel dual-path attention mechanism that integrates spatial and channel attention to mitigate microscopic feature attenuation and optimize gradient propagation.
- By analyzing WHO morphological criteria, we design acrosome anatomical constraint and head aspect ratio constraint, which are embedded into channel attention and hybrid loss functions respectively.
- We construct reversible feature-interaction by integrating reversible architectures with attention mechanisms, simultaneously preserving fine-grained microscopic features while optimizing GPU memory efficiency.
- We propose a hybrid loss function integrating corss-entropy loss, triplet loss and head aspect ratio constraint to improve discriminative capability for borderline samples, enhancing the model's interpretability and compliance with clinical standards.

II. Related Work

In this Section, we first review conventional deep learning approaches for sperm morphology analysis in Section II-A. Next, we examine advanced attention mechanisms in Section II-B. Finally, we discuss model optimization strategies and hybrid loss functions tailored for medical imaging tasks in Sections II-C and II-D, respectively, emphasizing the importance in improving interpretability and compliance with clinical standards.

A. Deep learning in sperm classification

Manual sperm detection is inefficient and highly subjective. To address this challenge, various approaches in deep learning have been explored for sperm classification.

Jabbari [10] introduces a hybrid capsule network and GAN architecture to improve classification. Similarly, Abbasi [6] applies GAN-based augmentation to balance datasets and enhance robustness. Yuzkat et al. [11] propose a soft voting fusion of six CNNs, achieving 85.18% accuracy on HuSHeM [49].

Liu et al. [8] use AlexNet-based transfer learning with 96.0% accuracy but rely on manual cropping and alignment, limiting automation and generalization. FT-VGG [12] reaches 94% accuracy with similar manual interventions. Ilhan et al. [15] achieve an accuracy of 85.42% using SURF features with an SVM classifier on their self-collected dataset, while a method based on DTCWT features attains 82.33% accuracy [16]. Ilhan et al. [17] also apply automatic directional masking with k-NN on HuSHeM, achieving 57.4% accuracy under five-fold CV after 93.5% correct masking—useful for automatic ROI extraction but with modest overall performance. Additionally, Ilhan et al. [18] adopt a two-stage network with soft voting, yielding 92.1% accuracy but complex training. Zhang et al. [19] employ a three-stage method requiring manual hyperparameter tuning, especially in pseudo-mask generation needing expert anatomical priors, reducing practicality.

Abbasi [20] presents a deep multi-task learning model combining transfer and multi-task learning, classifying sperm head, vacuoles, and acrosomes simultaneously with 84% accuracy on MHSMA. Ni et al. [21] develop a local constraint and label embedding multi-layer dictionary learning model with asymmetric Huber loss, validated on HuSHeM [49]. Yang et al. [22] propose a multi-object tracking algorithm integrating dynamic trajectory and morphological features, achieving 90.6% accuracy.

Though these studies show progress, most depend on manual interventions and do not fully utilize WHO clinical biological priors, limiting interpretability and automation. Inspired by [26], [32], [34], the WHO sperm standards are adopted as prior knowledge in this work, introducing acrosome anatomical constraint and head aspect ratio constraint to improve the model's adaptability to WHO standards. Therefore, our **MADRNet** significantly enhance accuracy.

B. Attention mechanism in microscopic images

In the field of microscopic image processing, attention mechanisms can dynamically localize key subtle structures, significantly improving feature discriminability while enhancing model interpretability. Ehsan et al. [30] propose Att Swin U-Net for skin lesions, which is an attention based extension of Swin U-Net used for medical image segmentation. They attempt to enhance the feature reusability by carefully designing skip connection paths. They believe that the classical cascade operation used in skip connection paths can be further improved by combining attention mechanisms. Shyam Lal et al. [31] propose NucleiSegNet—a robust deep

learning network architecture for nuclei segmentation in H&E-stained hepatic cancer histopathological images. The proposed architecture comprised three blocks: a robust residual block, a bottleneck block, and an attention decoder block. The attention decoder block employs a novel attention mechanism to effectively locate objects and enhance the proposed architecture's performance by reducing false positives. Chen et al. [36] propose the HAT model for low-level visual tasks such as image super-resolution, which integrates channel attention mechanism and window based local attention mechanism to explore deeper features. To prevent symptoms of pollen syndrome, Li et al. [37] construct an attention-based multi-scale feature fusion network based on light microscope images, which performs well in 8 pollen categories. Zhu et al. [38] propose a dual attention mechanism network for single image super-resolution. The result demonstrates significant improvements in both reconstructed image quality and radiologists' confidence for early-stage lung cancer diagnosis. However, most of the aforementioned methods use serialized spatial and channel attention, resulting in long paths that cause gradient decay and loss of microscopic features, thus limiting the capture of fine-grained morphological details.

In **MADRNet**, we leverage a parallel fusion of spatial and channel attention to implement a dual-path mechanism. The new approach effectively not only alleviates gradient vanishing, but also enhances the response to sperm acrosome and head microstructures.

C. Model optimization of deep learning

In the field of model optimization, various techniques have been proposed to reduce computational complexity without sacrificing performance. Li et al. [23] reduce the computational complexity of the ResNet101 model by 80% through channel pruning and layer fusion, resulting in only a 0.72% decrease in accuracy. Their approach allows the removal of 48 convolutional layers from the ResNet110 model without any noticeable loss in performance.

Jacobsen et al. [24] introduce reversible networks into the ResNet architecture, improving performance while reducing computational overhead. Similarly, Mangalam et al. [25] propose the Reversible Vision Transformer (RVT), a memory-efficient architecture designed for visual recognition tasks. RVT decouples GPU memory consumption from model depth, reducing memory usage to 1/15.5 of the original without compromising model complexity, parameter count, or accuracy.

Building on these advancements, Kashu et al. [29] apply the concept of reversible residual networks to 3D U-Net for volumetric segmentation. They develop partially and fully invertible versions of residual networks, significantly reducing memory usage while maintaining similar segmentation performance.

In **MADRNet**, we innovatively combine a parallel dual-path attention mechanism with a reversible architecture. This combination optimizes memory usage and enhances the retention of microscopic details, effectively mitigating information loss issues.

D. Hybrid loss function in medical model

Traditional loss functions demonstrate limited effectiveness in medical image segmentation tasks, particularly when addressing class imbalance, boundary ambiguity, and competing false positive/negative predictions. Therefore, the hybrid loss function plays a crucial role in the deep learning model. Chen et al. [39] propose hybrid strategies for different problems in their paper. Huang et al. [40] use a novel hybrid loss function with high sensitivity to boundaries to refine the image and obtain more accurate segmentation results. Tang at el. [41] achieve automatic segmentation and quantification of pneumothorax in CT images based on a hybrid loss attention mechanism. Liu et al. [8] combine a hybrid loss function with a deep supervised structure, effectively improving the segmentation performance of edge details, and applied it to the segmentation task of liver images. Yang et al. [44] use class balance loss functions to train whole heart analysis tasks. Many liver segmentation algorithms exhibit high sensitivity to ambiguous boundaries and heterogeneous pathologies, particularly under data scarcity. To address these challenges, Tan et al. [42] propose an automatic liver segmentation framework using a 3D convolutional neural network with a hybrid loss function. This approach combines cross-entropy loss, edge-preserving smoothness, and a shape constraint to improve segmentation accuracy, particularly at the boundaries and in capturing structural details.

Therefore, we also design a novel hybrid loss function that combines corss-entropy loss, triplet loss with head aspect ratio constraints. This design not only enhances the model's interpretability and compliance with clinical standards but also ensures high robustness on boundary samples.

III. Methodology

A. Overall design

Figure 1 illustrates the architecture of the proposed Morphology-Aware Dual-path Reversible Network (**MADRNet**), which comprises four core modules: backbone, reversible feature interaction, bilinear attention pooling, and classifier. The **MADRNet** architecture is formulated as:

$$\textbf{MADRNet}(x) = \mathcal{C}_{\text{classifier}}(\mathcal{P}_{\text{bilinear}}(\mathcal{R}^{(n)}_{\text{reversible}}(\mathcal{B}_{\text{backbone}}(x)))) \tag{1}$$

We choose ResNet50 as the low-level feature extractor in $\mathcal{B}_{\text{backbone}}$. The residual connections in ResNet50 effectively mitigate gradient vanishing issues [46], [47]. Also, it takes less memory consumption compared to other models (e.g., ResNet101) while maintaining comparable performance [46].

$\mathcal{R}^{(n)}_{\text{reversible}}$ is a reversible architecture with parallel dual-path attention mechanism, which considers acrosome anatomical constraint. This module aims to preserve detailed information while minimizing GPU resource utilization.

$\mathcal{P}_{\text{bilinear}}$ is a module to achieve bilinear pooling following the parallel dual-path attention component. It is used to capture co-variation relationships among feature channels, i.e., second-order statistics. Therefore, it can enhance the model's capability to discern subtle morphological differences in sperm cells.

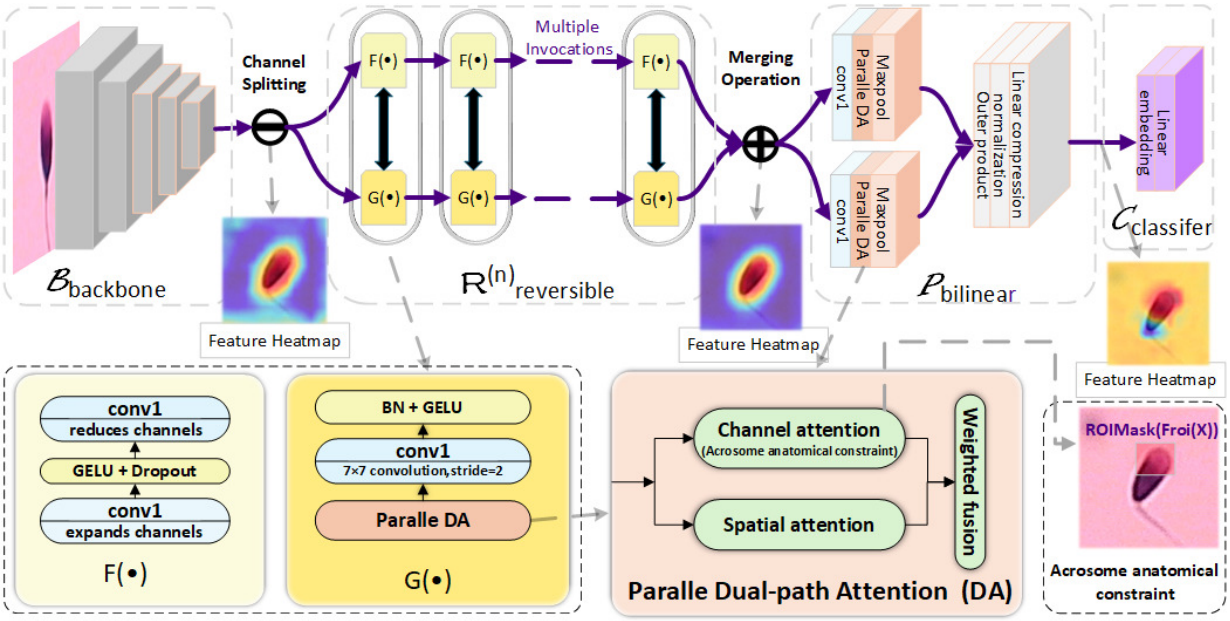


Fig. 1: Overall network architecture

$\mathcal{C}_{\text{classifier}}$ includes an embedding layer and an output layer. The embedding layer maps the pooled features into a low-dimensional space to form compact embeddings. These embeddings, which are in the low-dimensional space, play the role of input for a triplet loss function. This design enhances model robustness for boundary cases while ensuring compliance with WHO-defined sperm head aspect ratio.

In Section III-B, we will first elaborate the implementation details of the parallel dual-path attention (DA) mechanism, as it is utilized in both $\mathcal{R}^{(n)}_{\text{reversible}}$ and $\mathcal{P}_{\text{bilinear}}$. Then, the detailed design of $\mathcal{R}^{(n)}_{\text{reversible}}$ and the implementation of $\mathcal{P}_{\text{bilinear}}$ are be introduced in Section III-C and Section III-D, respectively. Additionally, Section III-E will provide a detailed explanation of the proposed hybrid morphology-aware loss function.

B. Parallel Dual-path Attention (DA) Mechanism

The parallel DA mechanism is designed to quantify weighted feature importance during the training process. Let the input feature be $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$, where C is the number of channels, H and W denote the height and width of the feature map. It includes two parts: channel attention and spatial attention, which output channel weight matrix and spatial weight matrix, respectively.

$$A_c = \sigma(W_c(\text{AVP}(\mathbf{X}))) \odot \text{ROIMask}(\mathcal{F}_{\text{roi}}(\mathbf{X})) \tag{2}$$

The **channel attention** operates using Equation (2). The input feature $\mathbf{X}$ is first processed by the $\mathcal{F}_{\text{roi}}(\cdot)$ function, where "roi" represents the region of interest that the model focuses on. The function consists of a global average pooling layer and a linear encoding layer, aiming to calculate the global channel response to reflect the degree of acrosome integrity. The ROIMask function further encodes the probability distribution of acrosome integrity and outputs a weight matrix for the acrosome region. Meanwhile, the feature $\mathbf{X}$ undergoes adaptive average pooling (AVP) and passes through W_c to obtain global channel weights matrix. W_c is the bottleneck structure to perform channel dimension reduction and expansion. Next, the output is compressed to the range [0,1] using the σ activation function. Finally, the A_c is obtained by multiplication ($\odot$) of global channel weight matrix and the acrosome region weight matrix.

$$A_s = \sigma(\text{Conv}_{3\times3}(\mathbf{X})) \tag{3}$$

The **spatial attention** operates as shown by Equation (3). A lightweight convolutional layer ($\text{Conv}_{3\times3}$) is utilized. It reduces the channel dimension to a single channel to avoid interference. σ activation function is applied on the single-channel dimension to generate A_s as the spatial weight matrix.

$$\mathbf{X}_{\text{out}} = \mathbf{X} \odot A_c + \mathbf{X} \odot A_s \tag{4}$$

After the channel and spatial attention processing, the A_c and A_s are fused in parallel as operated by Equation (4). Therefore, the proposed parallel DA attention mechanism enables adaptive calibration of channel and spatial feature mappings, effectively enhancing focus on subtle local features.

C. Reversible Feature Interaction $\mathcal{R}^{(n)}_{reversible}$

The reversible feature interaction module $\mathcal{R}^{(n)}_{\text{reversible}}$ is based on a reversible architecture, where each layer's operations can be exactly inverted during backpropagation. Therefore, the $\mathcal{R}^{(n)}_{\text{reversible}}$ module enables to mitigate information loss and preserve more detailed features without storing intermediate activations. Specifically, the module contains n executions of reversible feature interaction functions, F and G. The F function uses simple convolution and normalization to achieve shallow feature interaction. The G function mainly combines the parallel DA component described in Section III-B to enhance attention and a convolution component to fuse channel information. Through the combination, the interaction of deep

features is realized. Advantageously, the reversible architecture enables **MADRNet** to reduce GPU memory consumption.

Particularly in the forward propagation, before entering $\mathcal{R}^{(n)}_{\text{reversible}}$, feature $\mathbf{X}$ is split the channel dimensions into $\mathbf{X}_1$ and $\mathbf{X}_2$, as Equation (5) does.

$$\begin{cases} \mathbf{X}_1^{(l+1)} = \mathbf{X}_1^{(l)} + \mathsf{F}(\text{Normalization}(\mathbf{X}_2^{(l)})) \\ \mathbf{X}_2^{(l+1)} = \mathbf{X}_2^{(l)} + \mathsf{G}(\mathbf{X}_1^{(l+1)}) \end{cases} \tag{5}$$

where l (range $0 \sim n-1$) means the l-th execution in forward propagation. In the operation, $\mathbf{X}_2^{(l)}$ is normalized firstly and then fed into the F function. Then output is added to $\mathbf{X}_1^{(l)}$ to obtain $\mathbf{X}_1^{(l+1)}$. Simultaneously, $\mathbf{X}_1^{(l+1)}$ is input to the G function and the output is added to $\mathbf{X}_2^{(l)}$ to obtain $\mathbf{X}_2^{(l+1)}$. Moreover, F and G adopt a hierarchical parameter reuse mechanism, automatically inheriting the previous parameter configurations upon each invocation. This design preserves the independence of each layer through parameter passing while achieving semantic alignment across layers.

$$\begin{cases} \mathbf{X}_2^{(l)} = \mathbf{X}_2^{(l+1)} - \mathsf{G}\left(\mathbf{X}_1^{(l+1)}\right) \\ \mathbf{X}_1^{(l)} = \mathbf{X}_1^{(l+1)} - \mathsf{F}\left(\text{Normalization}(\mathbf{X}_2^{(l)})\right) \end{cases} \tag{6}$$

During backpropagation, the inverse computation of the forward propagation (as specified in Equation (6)) is consequently applied. $\mathbf{X}_2^{(l)}$ and $\mathbf{X}_1^{(l)}$ are successively calculated. This inversion enables lossless reconstruction of the model parameters while eliminating the need to store intermediate activation values. Therefore, the proposed reversible feature interaction module $\mathcal{R}^{(n)}_{\text{reversible}}$ preserves more detailed features and reduces memory usage.

D. Bilinear Attention Pooling $\mathcal{P}_{bilinear}$

To enhance discriminative capability for subtle morphological differences in sperm cells, a bilinear attention pooling module is devised $\mathcal{P}_{\text{bilinear}}$ to capture second-order statistics of features. Subsequent to $\mathcal{R}^{(n)}_{\text{reversible}}$, features $\mathbf{X_1}$ and $\mathbf{X_2}$ are merged into $\mathbf{X_{fused}}$ along the channel direction. The features $\mathbf{X_{fused}}$ are inputs for a pooling component, which is replicated in two branches. The component additionally incorporates Parallel DA mechanism (as introduced in Section III-B), which enhances the pooling process by selectively attending to both channel and spatial features. Equation (7) formally illustrates the pooling component.

$$\mathbf{B} = \text{GAP}(\text{Parallel DA}(\text{Conv}_{1\times1}(\mathbf{X_{fused}}))) \tag{7}$$

Where $\text{Conv}_{1\times1}$ denotes a 1×1 convolutional layer for channel dimensionality reduction, Parallel DA refers to the dual-path attention component described in Section III-B, and GAP denotes global average pooling.

With two independent output vectors $\mathbf{B_1}$ and $\mathbf{B_2}$ from the pooling component, outer product of the vectors are calculated to obtain $\mathbf{X_{OP}}$ as Equation (8) demonstrates.

$$\mathbf{X_{OP}} = \mathbf{B_1}^T\mathbf{B_2} \tag{8}$$

To capture the positive/negative correlations of features during the normalization process, signed square root normalization is applied to $\mathbf{X_{OP}}$, as shown in Equation (9).

$$\hat{\mathbf{X}} = \text{sign}(\mathbf{X_{OP}}) \odot \sqrt{|\mathbf{X_{OP}}| + \epsilon} \tag{9}$$

where, sign is the sign function, which is applied to retain the directional (positive or negative) information of the feature vectors. This prevents the square root operation from disrupting the original sign structure of $\mathbf{X_{OP}}$. $|\mathbf{X_{OP}}|$ is defined as the element-wise absolute value of $\mathbf{X_{OP}}$. This operation ensures that each element is suitable for square root computation while preserving the relative magnitude of values. ϵ is a small constant (set to 10^{-5}) introduced for numerical stability.

Following normalization, $\hat{\mathbf{X}}$ is passed through a fully connected transformation to capture the final second-order statistics, formulated in Equation (10).

$$\mathbf{X_{Output}} = \text{Linear}(\text{ReLU}(\text{BN}(\text{Linear}(\hat{\mathbf{X}})))). \tag{10}$$

where Linear denotes a fully connected layer, BN is batch normalization, and ReLU represents the rectified linear unit activation function. By capturing second-order statistics, richer structural information is provided to the model for classification. Therefore, the proposed bilinear attention pooling module enhances the model's capacity to distinguish subtle morphological differences, thereby improving the accuracy of sperm morphology classification.

E. Hybrid Morphology-Aware Loss Function $\mathcal{HL}$

We design a hybrid morphology-aware loss function $\mathcal{HL}$ to enhance the compatibility of model outputs with WHO sperm morphology standards and improve robustness for boundary samples. It consists of three components: the cross-entropy loss function (denoted as $\mathcal{L}_{CE}$), the triplet loss function ($\mathcal{L}_{Triplet}$), and the head aspect ratio constraint ($\mathcal{C}_{HARC}$). The definition of hybrid morphology-aware loss function is shown in Equation (11).

$$\mathcal{L}_{\mathcal{HL}} = \mathcal{L}_{CE} + \alpha\mathcal{L}_{Triplet} + \beta\mathcal{C}_{HARC} \tag{11}$$

The weights for the Triplet and HARC terms are learnable scalars, denoted as α and β, and are initialized to 0.4 and 0.3, respectively.

$\mathcal{L}_{CE}$ (defined by Equation (12)) is widely used in classification tasks. It measures the discrepancy between the predicted probability distribution and the true label distribution. By encouraging the model to assign high probabilities to the correct class, it provides the main supervisory signal that supports classification stability and class-discriminative learning.

$$\mathcal{L}_{CE} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{c=1}^{N} y_{i,c}\log(\hat{y}_{i,c}) \tag{12}$$

where B is the batch size, N is the number of classes, $y_{i,c}$ is the one-hot encoded ground truth label for sample i, and $\hat{y}_{i,c}$ is the predicted probability of sample i belonging to class c.

Given input embeddings, $\mathcal{L}_{Triplet}$ searches for anchor samples (denoted as a), positive samples (p) of the same class,

and negative samples (n) of different classes. It operates as shown in Equation (13).

$$\mathcal{L}_{Triplet} = \frac{1}{|T|} \sum_{(a,p,n)\in T} \max\left(d(a,p) - d(a,n) + m, 0\right) \tag{13}$$

where T is the set of valid triplets and $d(\cdot)$ denotes cosine similarity. $\mathcal{L}_{Triplet}$ compares the distances among anchor (a), positive (p), and negative (n) samples in the embedding space. It encourages the model to bring samples of the same class closer while pushing different-class samples apart, based on a margin m. This spatial separation helps the model learn more structured and discriminative representations.

$$\mathcal{C}_{HARC} = \frac{1}{B} \sum_{i=1}^{B} \max\left(|\gamma_i - 1.5| - \delta, 0\right) \tag{14}$$

$\mathcal{C}_{HARC}$ is formulated by Equation (14). Where γ_i denotes the aspect ratio of the bounding box surrounding the largest contour in the i-th sample, extracted using OpenCV. The reference value of 1.5 corresponds to the midpoint of the WHO-defined normal range, and $\delta = 0.3$ specifies the acceptable tolerance. This constraint imposes penalties on samples whose aspect ratios exceed the accepted range.

IV. Experiments

A. Experimental Setup

1) Datasets & Preprocessing: We evaluate the performance of MADRNet on the Human Sperm Head Morphology (HuSHeM) dataset [49]. The dataset comprises semen samples from fifteen patients at the Isfahan Fertility and Infertility Center. Following fixation and staining via the Diff-Quick protocol, sperm samples undergo imaging using an Olympus CX21 microscope equipped with a ×100 objective, ×10 eyepiece, and Sony SSC-DC58AP color camera. This process yields 725 raw images at 576×720 pixel resolution. Three clinical specialists manually crop sperm heads from these images and classify them into five morphological categories: Normal, Pyriform, Tapered, Amorphous, and Others. They retain only specimens achieving unanimous expert consensus. The final dataset contains 216 RGB sperm head images standardized to 131×131 pixels, distributed across four classes: Normal (54), Tapered (53), Pyriform (57), and Amorphous (52), as visualized in Figure 2. HuSHeM serves as an essential benchmark for developing computer-assisted sperm morphology classification models.

a Normal b Tapered c Pyriform d Amorphous

Fig. 2: Four morphological categories in HuSHeM dataset

The HuSHeM dataset provides class-wise image folders without patient-level identifiers; therefore, we adopt image-level stratified 5-fold cross-validation with a fixed random seed. The data split is performed prior to any augmentation. All images are resized to 512×512 pixels. Training employs lightweight augmentations, including random horizontal flipping and ±20° rotation, followed by ImageNet-style normalization. To preserve anatomical geometry, we avoid using perspective/shear transformations or synthetic masks. The head aspect ratio used by $\mathcal{C}_{HARC}$ is computed on the raw RGB images—before augmentation—based on the bounding rectangle of the largest external contour. For validation, a deterministic preprocessing pipeline is applied, consisting of resizing, center cropping, and normalization.

2) Evaluation Metrics: For the evaluation of our model's performance, we use four commonly applied metrics: accuracy, precision, recall, and F1 Score.

- **Accuracy** measures the overall correctness of the model by calculating the ratio of correct predictions to the total number of predictions. It is defined as:

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Samples}}$$

- **Precision** quantifies the model's ability to correctly identify positive samples. It is the ratio of true positive predictions to the total predicted positives:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

- **Recall**, also known as sensitivity, measures the model's ability to detect all positive samples. It is defined as the ratio of true positives to the total actual positives:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

- **F1 Score** is the harmonic mean of precision and recall, providing a single metric that balances both aspects. It is particularly useful in situations where there is a class imbalance:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

These metrics are essential for evaluating the performance of our model, as they provide a comprehensive assessment of both the positive and negative class predictions.

3) Implementation Details: All experiments are conducted using PyTorch on an NVIDIA RTX 4090 GPU.

We trained the model for 400 epochs using AdamW, adopting OneCycleLR to smoothly adjust the learning rate (lr). The $\mathcal{B}_{\text{backbone}}$ uses a low initial rate, while the $\mathcal{R}_{\text{reversible}}$ and $\mathcal{C}_{\text{classifier}}$ head use a high rate to accelerate adaptation. We repeated the F and G functions in our $\mathcal{R}_{\text{reversible}}$ three times. We set the total loss weight to 0.4, and applied gradient clipping to maintain training stability. Table I summarizes key hyperparameters.

B. Qualitative Visualization

We present two complementary qualitative visualizations: (i) an embedding-space representation using t-SNE, and (ii) pixel-level attention maps generated via Grad-CAM++.

We project the penultimate-layer embeddings into 2D using t-SNE and visualize the validation set, color-coded by class (perplexity = 15, initialization = PCA, random_state = 42;

TABLE I: Hyperparameters for the **MADRNet**

Hyperparameter	Value
Batch Size	64
Epochs	400
n	3
$lr_{backbone}$	3×10^{-4}
$lr_{reversible}$	1×10^{-3}
$lr_{classifier}$	1×10^{-3}
Optimizer	AdamW
Learning Rate Scheduler	OneCycleLR
Loss Weight Coefficient	0.4
Gradient Clipping (Max Norm)	1.0

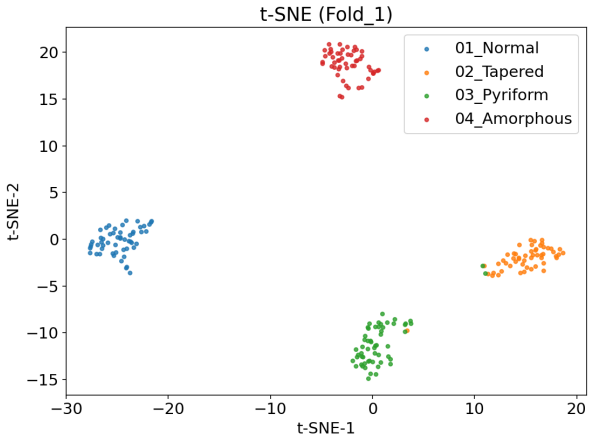


Fig. 3: t-SNE of penultimate-layer embeddings on the validation set (color-coded by class).

other parameters set to default). Figure 3 reveals clear inter-class separation and intra-class compactness, indicating that the learned representations are both morphology-consistent and discriminative. Additional folds are illustrated by Figure 12 in Appendix A.

To evaluate the clinical interpretability of **MADRNet**'s decision process, we apply Grad-CAM++ [50] to visualize which regions of the input images are most important. Figure 4 shows the results, where each representative sample of the four morphological classes are selected. In each group, the leftmost image represents the original input, followed by the attention heatmap obtained after extraction via $\mathcal{B}_{backbone}$. On the rightmost side, the feature heatmap is presented after undergoing $\mathcal{R}^{(n)}_{reversible}$ and $\mathcal{P}_{bilinear}$.

By comparing the heatmap distributions across sample groups, it can be observed that module $\mathcal{B}_{backbone}$ exhibits significant heterogeneity in its focus on discriminative features. After optimization with module $\mathcal{R}^{(n)}_{reversible}$ and module $\mathcal{P}_{bilinear}$, the salient regions of heatmaps consistently converge on the acrosomal structure. This observation validates the effectiveness of acrosomal anatomical constraint within the parallel DA mechanism.

C. Performance and Comparison

We evaluate the proposed **MADRNet** on the HuSHeM dataset using a five-fold cross-validation protocol. As shown in

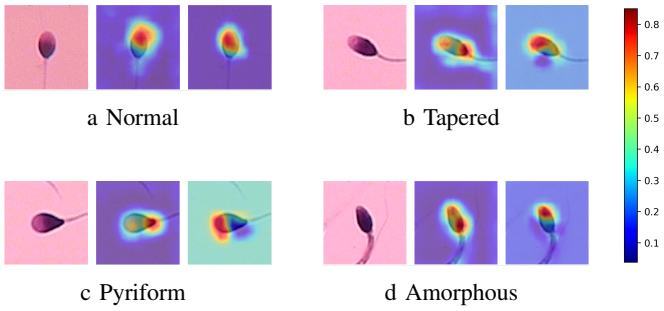


Fig. 4: Class-specific attention heatmaps demonstrating **MADRNet**'s interpretability and overall performance.

Figure 5, the training process converges at about 340 epochs, ultimately achieving 96.3% accuracy, 96.6% precision, 96.8% recall, and a 96.8% F1 score. These results underscore the robustness and efficiency of our approach in sperm morphology classification tasks.

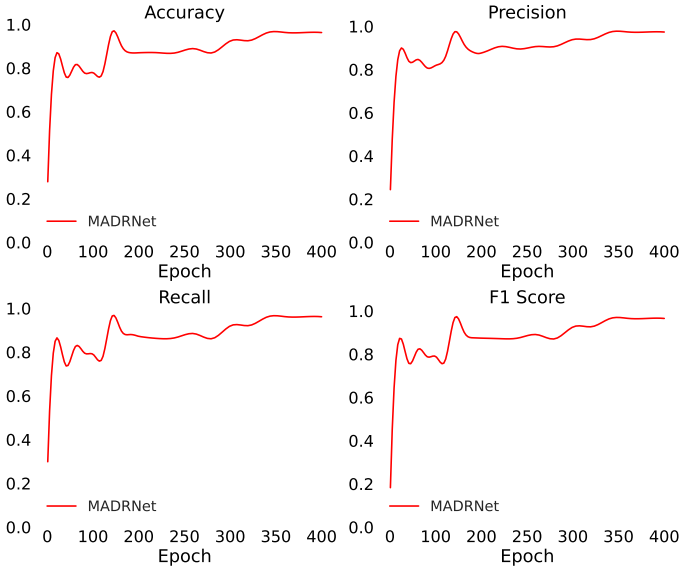


Fig. 5: Performance of **MADRNet** on the HuSHeM validation set

Table II summarize the comparison result with related works, indicating that our model exhibits superior performance. We report the mean ± standard deviation across five-fold cross-validation, along with 95% confidence intervals computed using the t-distribution (degrees of freedom (df) = 4). In the column of Pre-training, IN denotes ImageNet pre-trained weights, US denotes unsupervised pre-training, and SMIDS denotes pre-training on the SMIDS dataset. TL [8], FT-VGG [12] and Zhang et al. [19] achieve high accuracy, but they rely on human intervention, such as manual cropping and head alignment. Manual intervention limits the model's autonomy, making it impractical in real world applications. In contrast, our model achieves an accuracy rate of 96.3%, without any manual intervention.

The inference time for each image is 32 ms, and the model parameters are 44.35M, which is 1/3 of FT-VGG [12]. Meanwhile, the lightweight version parameter is only 12.34M,

TABLE II: Comparison of different methods on HuSHeM dataset.

Method	Pre-training	No Human Intervention	Accuracy(%)	Precision(%)	Recall(%)	F1 Score(%)
Yuzkat et al. [11]	×	✓	85.2	85.2	85.3	85.3
Ilhan et al. (2019) [17]	×	✓	57.4	–	–	–
Ilhan et al. (2020) [14]	×	✓	86.6	–	–	–
Ilhan et al. (2022) [18]	IN+SMIDS	✓	92.1	92.3	92.1	92.2
FT-VGG [12]	IN	×	94.1	94.3	94.1	94.2
TL [8]	IN	×	96.4	96.4	96.4	96.4
Zhang et al. [19]	IN+US	×	96.5	96.8	96.6	96.6
MADRNet (ours)	IN	✓	**96.3 ± 0.30**†	**96.6 ± 0.32**†	**96.8 ± 0.29**†	**96.8 ± 0.37**†

† mean ± standard deviation over five folds with 95% confidence intervals based on the t-distribution (df = 4).
– Not reported in the original paper.

which is only 1/11 of the FT-VGG model [12]. We will discuss the details of lightweight version in Section IV-F.

To thoroughly investigate the performance of the proposed model across different categories, we conduct a detailed analysis of the confusion matrix, as shown in Figure 6. Among all categories, the prediction accuracy for normal sperm cells is the highest, reaching 99%. The average prediction accuracy for tapered and pyriform sperm cells is relatively lower, at 94.3% and 95.1%, respectively. The high accuracy for normal sperm cells can primarily attribute to the consistency of their contours, which makes them easy to distinguish from other categories. However, due to the high morphological similarity between tapered and pyriform sperm cells, the model often faces challenges in differentiating between these two types. Thus, the misclassification rate is about 4.8% (or 4.2%) for incorrectly identifying tapered (or pyriform) as pyriform (or tapered) sperm cells. As for amorphous sperm cells, their irregular contours allowed them to be easily distinguished from other categories, resulting in a correct prediction rate of 99%.

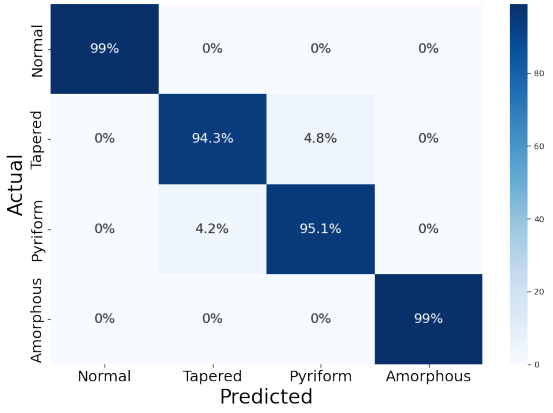


Fig. 6: Confusion matrix for four categories

In addition to the confusion matrix, Table III presents per-class Accuracy, Precision, Recall, and F1 scores, averaged over five folds. Consistent with Figure 6, the Normal and Amorphous classes achieve the highest performance across all metrics (approximately 0.99), while the Tapered and Pyriform classes show slightly lower scores (around 0.94–0.95), reflecting their morphological similarity. It is worth noting that since Figure 6 is row-normalized, the per-class Accuracy column (one-vs-all accuracy) numerically corresponds to the diagonal Recall values in this setting.

TABLE III: Per-class performance (5-fold).

Class	Accuracy	Precision	Recall	F1
Normal	0.990	0.988	0.990	0.989
Tapered	0.943	0.940	0.943	0.943
Pyriform	0.951	0.950	0.951	0.951
Amorphous	0.990	0.988	0.990	0.989

Furthermore, we compare the performance of the **MADRNet** model with FT-VGG [12], Zhang et al. [19], as shown in Figure 7. Compared to FT-VGG, **MADRNet** exhibits a decline only in the tapered classification, while it shows improvements in the other three categories. When compared to Zhang et al. [19], **MADRNet** maintains similar accuracy in the normal classification, with a slight decrease in the tapered classification. However, **MADRNet** demonstrates significant enhance in both pyriform and amorphous classifications, particularly in the Amorphous category.

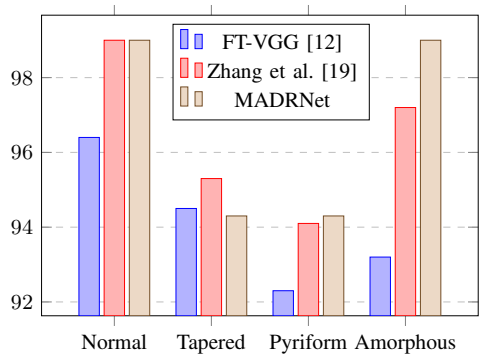


Fig. 7: Accuracy in four categories

D. Ablation Study

1) Effectiveness of Proposed Modules: To comprehensively evaluate each proposed modules and components, we conduct ablation study on $\mathcal{R}^{(n)}_{\text{reversible}}$ (refer to Section III-C), $\mathcal{P}_{\text{bilinear}}$ (refer to Section III-D), and $\mathcal{HL}$ (refer to Section III-E), respectively. Note that the parallel DA component (refer to

Section III-B) is employed by module $\mathcal{R}^{(n)}_{\text{reversible}}$ and module $\mathcal{P}_{\text{bilinear}}$, thus the ablation experiments do not include the component. The ablation configurations are structured as follows:

1) baseline: The baseline model uses the ResNet50 $\mathcal{B}_{\text{backbone}}$ without any proposed modules or components.
2) + $\mathcal{R}^{(n)}_{\text{reversible}}$: Extends the ResNet50 baseline by incorporating the reversible feature interaction module.
3) ++ $\mathcal{P}_{\text{bilinear}}$: Further adds the bilinear attention pooling module to the above configuration.
4) +++ $\mathcal{HL}$: Further incorporates the hybrid morphology-aware loss function to address class imbalance and enhance feature discriminability.

TABLE IV: Ablation Study Results

baseline	$\mathcal{R}^{(n)}_{\text{reversible}}$	$\mathcal{P}_{\text{bilinear}}$	$\mathcal{HL}$	Accuracy	F1 Score
✓				79.2	78.5
✓	✓			81.5	83.7
✓	✓	✓		89.1	90.4
✓	✓	✓	✓	96.3	96.8

Table IV shows the results of different configurations. The baseline model achieves an accuracy of 79.2% and an F1 score of 78.5%. On this basis, $\mathcal{R}^{(n)}_{\text{reversible}}$, and $\mathcal{P}_{\text{bilinear}}$ are gradually added, achieving accuracy increases of 9.3% and 7.6% respectively. After joining the $\mathcal{HL}$, the accuracy increases by 7.2%.

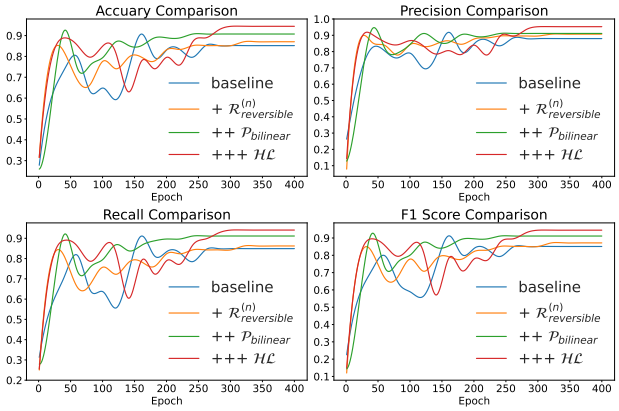


Fig. 8: Comparison of ablation experiments

Figure 8 presents the trends of four evaluation metrics across training epochs under different ablation configurations. The baseline model (blue) shows limited learning capacity, with all metrics plateauing below 0.80 and substantial instability throughout training. Upon introducing the $\mathcal{R}^{(n)}_{\text{reversible}}$ (green), noticeable performance improvements emerge, particularly within the first 100 epochs. However, convergence is only observed after approximately 200 epochs, indicating delayed stabilization despite enhanced feature preservation.

With the addition of the $\mathcal{P}_{\text{bilinear}}$ (orange), the model achieves faster and more consistent improvements, especially in precision and F1 score. Nonetheless, full convergence remains gradual and is not reached until around epoch 250. Finally, the complete configuration incorporating the $\mathcal{HL}$ (red) exhibits rapid early-stage gains but also notable mid-stage fluctuations, particularly between epochs 80 and 150. These are attributed to the dynamic weighting of triplet loss and head aspect ratio constraints. The model ultimately stabilizes after epoch 250, achieving consistently high scores across all metrics.

Overall, Figure 8 illustrates that each proposed module contributes incrementally to model performance. The full **MADRNet** configuration not only delivers superior accuracy and robustness, but also demonstrates a more structured optimization trajectory.

2) Effectiveness of Proposed Constraints: In the experimental stage, we also examine the contributions of two morphological constraints to model performance. One is the acrosome anatomical constraint (AAC), which is integrated into the parallel DA component. The other is the head aspect ratio constraint (HARC), which is incorporated into the hybrid morphology-aware loss. We conduct experiments on the HuSHeM dataset after pruning and combining the two constraints, with the results presented in Table V. The removal of the two constraints cause the model accuracy to drop to 91.7%. Starting from this baseline, the reintroduction of AAC and HARC single-handedly improve the accuracy by 2.9% and 1.8%, respectively. These experimental results confirm the effectiveness of the proposed constraints.

TABLE V: Ablation Study Results

AAC	HARC	Accuracy	F1 Score
		91.7	91
✓		94.6	93.3
	✓	93.5	94.1
✓	✓	96.3	96.8

E. External Verification

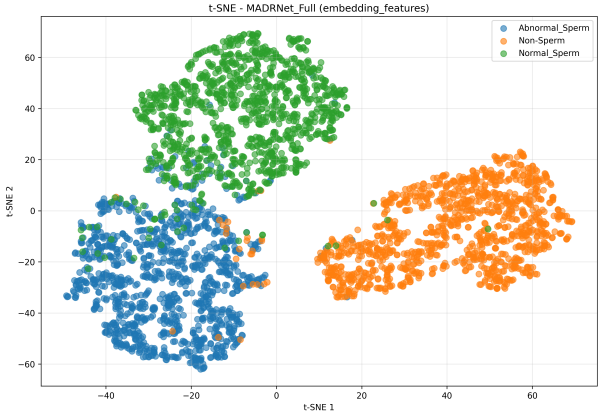
To validate the generalizability of MADRNet, we conduct an independent evaluation on the larger and clinically distinct SMIDS dataset [13]. As summarized in Table VI, our model achieves a competitive accuracy of 89.33% with a narrow 95% confidence interval of 88.01% to 90.66%, closely approaching the performance of leading methods [11], [18] on this benchmark. This demonstrates that MADRNet's architectural advantages generalize effectively beyond its training data.

TABLE VI: Comparison on the SMIDS dataset (Accuracy only, 5-fold cross-validation).

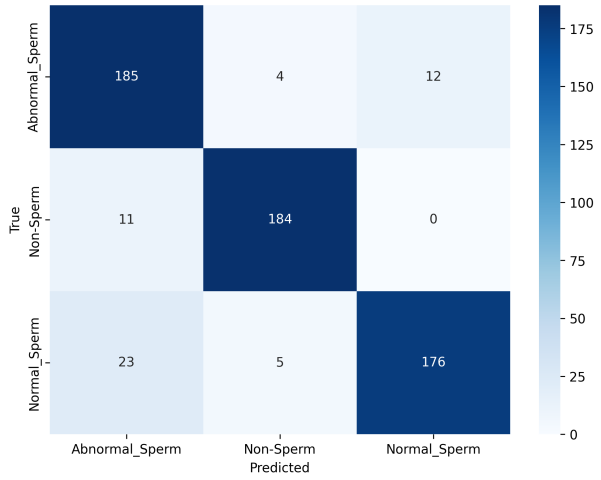
Model	**Accuracy (%)**
VGG19 [13]	87.00
MobileNetV1 [52]	88.00
MobileNetV2 [52]	87.00
Yüzkat et al. [11]	90.73
Ilhan et al.(2020) [14]	85.70
Ilhan et al.(2022) [18]	90.87
Ours (MADRNet)	**89.33±1.07 (95% CI 88.01–90.66)**

Values for prior work are point estimates reported in their papers. Our result is the mean±sd over 5 folds with a two-sided 95% confidence interval (df=4, $t = 2.776$).

Qualitative results further reinforce MADRNet's robustness. The clear inter-class separation in the t-SNE plot (Figure 9a) and the pronounced diagonal in the confusion matrix (Figure 9b) indicate that the model effectively learns discriminative, semantically meaningful features on this external dataset. The primary confusion arises between the morphologically similar 'Abnormal sperm' and 'Normal sperm' classes, which aligns with clinical observations.



a t-SNE of validation embeddings.



b Confusion matrix (5-fold average).

Fig. 9: SMIDS external replication under the unified protocol (same input size/augmentation; no resampling or manual cropping).

In summary, the competitive performance and structured feature representations on SMIDS demonstrate MADRNet's robustness and generalization, highlighting its potential across diverse clinical settings.

F. Model Lightweighting

The reversible feature interaction module $\mathcal{R}^{(n)}_{\text{reversible}}$ not only enables to enhance performance, but also facilitates to make the model lightweight. We test the performance through finetuning the parameter n, which represents the execution times of F and G algorithms in Equation (5). F and G adopt a hierarchical parameter reuse mechanism. Thus, n do not affect the total parameter count, but influence GPU memory occupation and inference speed. The results are presented in Figure 10. $n = 0$ indicates that the $\mathcal{R}^{(n)}_{\text{reversible}}$ module is not used.

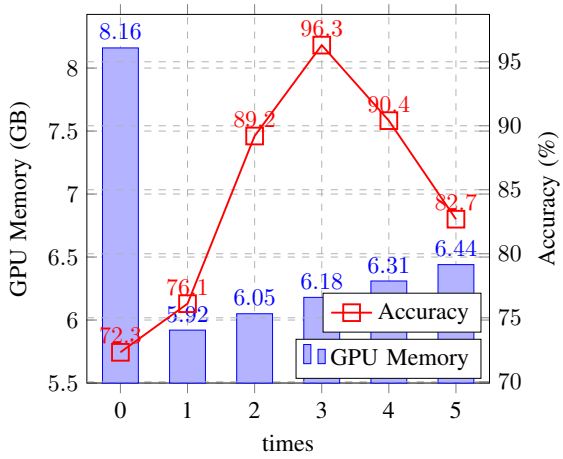


Fig. 10: Performance comparison of different execution times of F and G in $\mathcal{R}^{(n)}_{\text{reversible}}$

It can be seen that GPU memory consumption increases by about 0.13 GB for every increment of n, given $n > 0$. However, increasing n does not yield linear accuracy improvements. The accuracy improves significantly when $n \leq 3$, but degrades when $n > 3$. That is primarily because increasing n leads to more feature fusion or interaction, which consequently enables high accuracy. However, too much feature fusion or interaction may cause gradient instability, resulting in low accuracy. It can be observed that the memory usage decreases by 24.3% when $n = 3$ compared to $n = 0$. Performance testing demonstrates an inference latency of 32ms per image, which is promising for real-time image analysis.

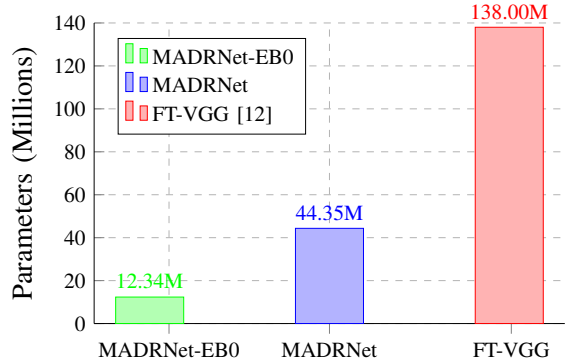


Fig. 11: Parameter comparison of different models

Figure 11 presents the number of parameters of FT-VGG [12] and the proposed **MADRNet**. It can be seen that our **MADRNet** has a total parameter count of 44.35M, equivalent to 1/3 of FT-VGG. To further simplify the model, we replaced the backbone with EfficientNet-B0 to obtain **MADRNet-EB0**. EfficientNet-B0 is close to the accuracy of resnet50 and has fewer model parameters. Consequently, **MADRNet-EB0** has only 12.34M parameters but with 93% accuracy. This modification significantly reduces hardware requirements

for terminal medical equipment, especially for organizations equipped with mid-range GPUs with <16GB VRAM.

To address the accuracy degradation of the lightweight model **MADRNet-EB0**, a hierarchical knowledge transfer approach based on a teacher-student framework [51] can be employed. Particularly, **MADRNet** serves as the teacher network for **MADRNet-EB0**, which is as the student network. That will be future work and the details are omit here.

V. Discussion

A. Model Innovation

In this study, we propose the **MADRNet**, which integrates a dual-path attention mechanism and a reversible architecture for sperm classification. The key innovation lies in the incorporation of both parallel spatial and channel attention, alongside acrosome anatomical and head aspect ratio constraints, which directly align with WHO standards. These features not only enhance the model's accuracy but also improve its clinical interpretability, enabling precise sperm analysis while optimizing GPU memory usage. This makes **MADRNet** a promising solution for automated sperm morphology classification.

B. Consideration for Transformer-based Models

While transformer architectures have shown strong performance in computer vision, we opt not to incorporate them into **MADRNet** due to the limited size of the HuSHeM dataset. Instead, **MADRNet**, built on CNNs, offers superior data efficiency by leveraging spatial inductive biases and domain-specific constraints. This design achieves competitive accuracy without the extensive pre-training data typically required by transformer models, making it more suitable for HuSHeM's limited annotations. Nonetheless, we remain interested in exploring transformer-based models on the larger SMIDS dataset and leave this as future work.

C. Limitations and Future Work

While **MADRNet** achieves high accuracy, it still faces challenges in directly generalizing to other datasets. Future work will focus on applying the model to clinical stained sperm data, exploring its accuracy in real-world settings. Additionally, we plan to develop a model specifically designed for unstained sperm datasets, aiming to reduce reliance on lab processes and further enhance clinical efficiency.

VI. Conclusion

In this paper, we propose **MADRNet**, designed to recognize human sperm morphology while aligning with WHO standards and maintaining computational efficiency. The **MADRNet** mainly contains four modules. First, $\mathcal{B}_{\text{backbone}}$ uses ResNet50 to extract low-level features from the input images. Second, $\mathcal{R}^{(n)}_{\text{reversible}}$ preserves fine-grained details while reducing memory consumption. Third, $\mathcal{P}_{\text{bilinear}}$ captures second-order statistics across channels and spatial dimensions, thereby enhancing the network's ability to distinguish subtle morphological differences. Finally, $\mathcal{C}_{\text{classifier}}$ applies $\mathcal{HL}$—combining cross-entropy, triplet loss, and head aspect ratio constraint—to optimize decision boundaries for hard samples. Ablation studies have demonstrated the show significant synergistic effect of each module. The experimental results show that **MADRNet** achieves 96.3% classification accuracy and 96.8% F1 score. The an average inference latency of just 32ms per image. Besides, it is lightweight with only 44.35M model parameters. In the future, we plan to extend our approach to other cell-morphology classification tasks to highlight broader applicability in medical image analysis.

[1] World Health Organization. *WHO Laboratory Manual for the Examination and Processing of Human Semen*, 6th ed. Geneva, 2021. Online: https://www.who.int/publications/i/item/9789240030787

[2] V. Chang, A. Garcia, N. Hitschfeld and S. Härtel, "Gold-Standard for Computer-Assisted Morphological Sperm Analysis," *Computers in Biology and Medicine*, 2017, 83: 143-150.

[3] Y. Michailov, L. Nemerovsky, Y. Ghetler, M. Finkelstein, O. Schonberger, A. Wiser, A. Raziel, B. Saar-Ryss, I. Ben-Ami, I. Ben-Ami, O. Kaplanski, N. and Miller, "Stain-Free Sperm Analysis and Selection for Intracytoplasmic Sperm Injection Complying with WHO Strict Normal Criteria" *Biomedicines*, 2023, 11(10): 2614.

[4] Y. Wang, J. Yang, Y. Jia, C. Xiong, T. Meng, H. Guan, W. Xia, M. Ding and M. Yuchi, "Variability in the Morphologic Assessment of Human Sperm: Use of the Strict Criteria Recommended by the World Health Organization in 2010," *Fertility and Sterility*, 2014, 101(4): 945-949.

[5] R. Marín and V. Chang, "Impact of Transfer Learning for Human Sperm Segmentation Using Deep Learning," *Computers in Biology and Medicine*, 2021, 136: 104687.

[6] A. Abbasi, S. Bahrami, T. Hemmati and S.A. Mirroshandel, "Transfer-GAN: Data Augmentation Using a Fine-Tuned GAN for Sperm Morphology Classification," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 2023, 11: 2440-2456.

[7] Y. Cai, Y. Zhou, Q. Han, J. Sun, X. Kong, J. Li and X. Zhang, Reversible column networks. arXiv preprint arXiv:2212.11696, 2022.

[8] R. Liu, M. Wang, M. Wang, J. Yin, Y. Yuan and J. Liu, "Automatic Microscopy Analysis with Transfer Learning for Classification of Human Sperm," *Applied Sciences*, 2021, 11(12):

[9] N. Sapkota, Y. Zhang, S. Li, P. Liang, Z. Zhao, J. Zhang, X. Zha, Y. Zhou, Y. Cao and D. Z. Chen, "SHMC-Net: A Mask-Guided Feature Fusion Network for Sperm Head Morphology Classification," *IEEE International Symposium on Biomedical Imaging*, 2024, 1-5.

[10] H. Jabbari and N. Bigdeli, "New Conditional Generative Adversarial Capsule Network for Imbalanced Classification of Human Sperm Head Images," *Neural Computing and Applications*, 2023, 35(27): 19919-19934.

[11] M. Yüzkat, H.O. Ilhan and N. Aydin, "Multi-Model CNN Fusion for Sperm Morphology Analysis," *Computers in Biology and Medicine*, 2021, 137: 104790.

[12] J. Riordon, C. McCallum and D. Sinton, "Deep Learning for the Classification of Human Sperm," *Computers in Biology and Medicine*, 2019, 111: 103342.

[13] H. O. Ilhan, I. O. Sigirci, G. Serbes and N. Aydin, "A fully automated hybrid human sperm detection and classification system based on mobilenet and the performance comparison with conventional methods," *Medical & Biological Engineering & Computing*, 2020, 58: 1047–1068.

[14] H. O. Ilhan, G. Serbes and N. Aydin, "Automated Sperm Morphology Analysis Approach Using a Directional Masking Technique," *Computers in Biology and Medicine*, 2020, 122: 103845.

[15] H. O. Ilhan, G. Serbes and N. Aydin, "Dual Tree Complex Wavelet Transform Based Sperm Abnormality Classification," *Proc. 41st Int. Conf. Telecommun. Signal Process. (TSP)*, 2018, 578-580.

[16] H. O. Ilhan, I. O. Sigirci, G. Serbes and N. Aydin, "The Effect of Nonlinear Wavelet Transform Based De-noising in Sperm Abnormality Classification," *2018 3rd Int. Conf. Comput. Sci. Eng. (UBMK)*, Sarajevo, Bosnia and Herzegovina, 2018.

[17] H. O. Ilhan, G. Serbes and N. Aydin, "Automatic Directional Masking Technique for Better Sperm Morphology Segmentation and Classification Analysis," *Electron. Lett.*, 2019, 55: 256-258.

[18] H.O. Ilhan and G. Serbes, "Sperm Morphology Analysis by Using the Fusion of Two-Stage Fine-Tuned Deep Networks," *Biomedical Signal Processing and Control*, 2022, 71: 103246.

[19] Y. Zhang, J. Zhang, X. Zha, Y. Zhou, Y. Cao and D. Chen, "Improving Human Sperm Head Morphology Classification with Unsupervised Anatomical Feature Distillation," *IEEE International Symposium on Biomedical Imaging*, 2022, 1-5.

[20] A. Abbasi, E. Miahi and S.A. Mirroshandel, "Effect of Deep Transfer and Multi-Task Learning on Sperm Abnormality Detection," *Computers in Biology and Medicine*, 2021, 128: 104121.

[21] T. Ni, Y. Ding, J. Xue, K. Xia, X. Gu and Y. Jiang, "Local Constraint and Label Embedding Multi-Layer Dictionary Learning for Sperm Head Classification," *ACM Transactions on Multimedia Computing Communications and Applications*, 2021, 17(3s): 1-16.

[22] H. Yang, M. Ma, X. Chen, G. Chen, Y. Shen, L. Zhao, J. Wang, F. Yan, D. Huang et al., "Multidimensional Morphological Analysis of Live Sperm Based on Multiple-Target Tracking," *Computational and Structural Biotechnology Journal*, 2024, 24: 176-184.

[23] Q. Li, H. Li and L. Meng, "Deep Learning Architecture Improvement Based on Dynamic Pruning and Layer Fusion," *Electronics*, 2023, 12(5): 1208.

[24] J. Behrmann, W. Grathwohl, R.T.Q. Chen, D. Duvenaud and J.-H. Jacobsen, "Invertible Residual Networks," *International Conference on Machine Learning*, 2019: 573-582.

[25] K. Mangalam, H. Fan, Y. Li, C.-Y. Wu, B. Xiong, C. Feichtenhofer and J. Malik, "Reversible Vision Transformers," *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, 10830-10840.

[26] X. Li, L. Li and K. Zhang, "DPKI-Net: Dual Prior Knowledge Injection Network for Multitask 3-D Medical Image Segmentation and Landmark Localization," *IEEE Transactions on Instrumentation and Measurement*, 2025, 74: 1-12.

[27] Z. Zhang, B. Qi, S. Ou and C. Shi,"Real-Time Sperm Detection Using Lightweight YOLOv5," *2022 IEEE 8th International Conference on Computer and Communications (ICCC)*, 2022, 1829-1834.

[28] S. Kosuge and T. Hamagami, "Sperm Detection and Tracking Model Using HDE Transformer With Spatio-Temporal Deformable Attention for Sperm Analysis Automation," *IEEE Access*, 2025, 51978-51985.

[29] K. Yamazaki, V.S. Rathour and T.H.N. Le, "Invertible Residual Network with Regularization for Effective Volumetric Segmentation," *Medical Imaging 2021: Image Processing*, 2022, 11596: 269-275.

[30] E.K. Aghdam, R. Azad, M. Zarvani and D. Merhof, "Attention Swin U-Net: Cross-Contextual Attention Mechanism for Skin Lesion Segmentation," *IEEE 20th International Symposium on Biomedical Imaging*, 2023, 1-5.

[31] S. Lal, D. Das, K. Alabhya, A. Kanfade, A. Kumar and J. Kini, "NucleiSegNet: Robust Deep Learning Architecture for the Nuclei Segmentation of Liver Cancer Histopathology Images," *Computers in Biology and Medicine*, 2021, 128:104075.

[32] X. You, J. He, J. Yang and Y. Gu, "Learning With Explicit Shape Priors for Medical Image Segmentation," *IEEE Transactions on Medical Imaging*, 2025, 2(44): 927-940.

[33] J. Zhang, X. Zhang, Y. Xu and W. Chen, "Spatial Prior-Embedded Neural Networks for Medical Image Segmentation," *Journal of Computer-Aided Design & Computer Graphics*, 2021, 33(8): 1287-1294.

[34] Y. Zhang, N. Meng, M. Zhao and T. Zhang,"RASpine: Regional Attention Lateral Spinal Segmentation based on Anatomical Prior Knowledge," *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2024, 1-4.

[35] V. Chang, "Segmentation and Classification of Human Sperm Heads Towards Morphological Sperm Analysis," *Universidad de Chile*, Santiago, Chile, 2015.

[36] X. Chen, X. Wang, J. Zhou, Y. Qiao and C. Dong, "Activating More Pixels in Image Super-Resolution Transformer," *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, 22367-22377.

[37] J. Li, Q. Wang, C. Xiong, L. Zhao, W. Cheng and X. Xu, "AMFF-Net: An Attention-Based Multi-Scale Feature Fusion Network for Allergic Pollen Detection," *Expert Systems with Applications*, 2024, 235: 121158.

[38] D. Zhu, D. Sun and D. Wang, "Dual Attention Mechanism Network for Lung Cancer Images Super-Resolution," *Computer Methods and Programs in Biomedicine*, 2022, 226: 107101.

[39] Y. Chen, W. Zhang, H. Lin, C. Zheng, T. Zhou, L. Feng, Z. Yi and L. Liu, "A Survey of Loss Functions of Medical Image Segmentation Algorithms," *Journal of Biomedical Engineering*, 2023, 40(2): 392-400.

[40] Y. Huang, Z. Shi, Z. Wang and Z. Wang, "Improved U-Net Liver Medical Image Segmentation Method Based on Hybrid Loss Function," *Laser & Optoelectronics Progress*, 2020, 57(22): 221003.

[41] Q. Tang, Z. Liu, Q. Wang, J. Huang, B. Xue and Y. Zhou, "Automatic Segmentation and Quantification of CT Pneumothorax Based on Hybrid Loss Attention," *Computer Applications and Software*, 2024, 41(12): 214-222.

[42] M. Tan, F. Wu, D. Kong and X. Mao, "Automatic Liver Segmentation Using 3D Convolutional Neural Networks With Hybrid Loss Function," *Medical Physics*, 2021, 48(4): 1707-1719.

[43] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, "Attention Is All You Need," *Advances in Neural Information Processing Systems*, 2017, 30.

[44] X. Yang, C. Bian, L. Yu, D. Ni and P.A. Heng, "Hybrid Loss Guided Convolutional Networks for Whole Heart Parsing," *Statistical Atlases and Computational Models of the Heart: ACDC and MMWHS Challenges*, 2018, 215-223.

[45] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *arXiv preprint arXiv:2010.11929*, 2020.

[46] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2016.

[47] H. Yu, T. Hou, D. Shen, Y. Liu, H. Liu and Q. Du, "Medical Equipment Detection Based on ResNet-50 Deep Learning Model," *Proc. Int. Conf. Intell. Syst. Comput. Netw.*, 2025, 1-7.

[48] Y. Cai, Y. Zhou, Q. Han, J. Sun, X. Kong, J. Li and X. Zhang, "Reversible Column Networks," *arXiv preprint arXiv:2212.11696*, 2023.

[49] F. Shaker, S.A. Monadjemi, J. Alirezaie, and A.R. Naghsh-Nilchi, "A dictionary learning approach for human sperm heads classification," *Computers in Biology and Medicine*, 2017, 91:181–190.

[50] A. Chattopadhay, A. Sarkar, P. Howlader and V. N. Balasubramanian, "Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks," *IEEE Winter Conference on Applications of Computer Vision*, 2018, 839-847.

[51] G. Hinton, O. Vinyals and J. Dean, "Distilling the knowledge in a neural network[J],"*arXiv preprint arXiv:1503.02531*,2015.

[52] O. L. Tortumlu and H. O. Ilhan, "The analysis of mobile platform based CNN networks in the classification of sperm morphology," *2020 Medical Technologies Congress (TIPTEKNO)*, 2020: 1-4.

[53] A. Aktas, G. Serbes, M. H. Yigit, N. Aydin, H. Uzun and H. O. Ilhan, "Hi-LabSpermMorpho: A Novel Expert-Labeled Dataset With Extensive Abnormality Classes for Deep Learning-Based Sperm Morphology Analysis," *IEEE Access*, 2024, 12: 196070-196091.

APPENDIX

A. t-SNE Embedding Visualization of HuSHeM

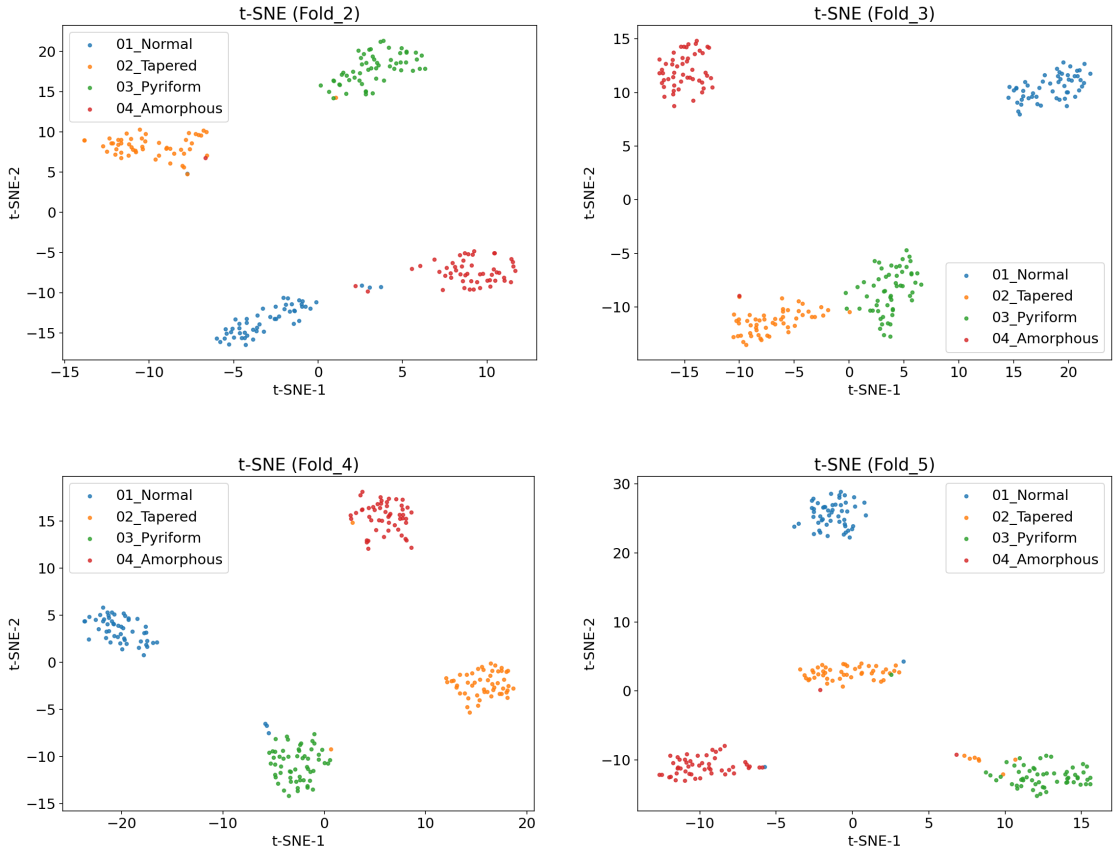


Fig. 12: t-SNE Embedding Visualization of HuSHeM: Visualization of high-dimensional embedding vectors projected into 2D space using t-SNE dimensionality reduction technique reveals sample distributions across different data folds. Different colors represent distinct class labels, allowing intuitive observation of clustering effectiveness and separation degree between categories in the embedding space.