Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 7 (2025)

Sperm Morphology Classification Using Xception-CBAM: A Deep Learning Approach on the SMIDS Dataset

[1]Ibadullah, [2]Muhammad Salih Tanveer , [3]Murad Khan , [4]Hayat Ur Rahman

Article Details

Key Words: Sperm Morphology Classification, Medical Image Analysis, Xception Network, CBAM Attention Mechanism, Deep Learning, SMIDS Dataset, Male Infertility Diagnosis, Computer-Aided Diagnosis

Ibadullah
Department of Computing, Riphah International University, Islamabad, Pakistan. Corresponding Author Email: ibad.ullah@riphah.edu.pk
Muhammad Salih Tanveer
COMSATS University, Islamabad, Pakistan. salih.hashmi.1999@gmail.com
Murad Khan
COMSATS University, Islamabad, Pakistan. muradkhanuswat@gmail.com
Hayat Ur Rahman
Department of MLT Riphah International University Malakand.
Hayat.rahman@riphah.edu.pk

ABSTRACT

Accurate classification of sperm morphology is fundamental in evaluating male fertility and supporting reproductive health diagnostics. However, conventional manual assessment methods are often subjective, labor-intensive, and prone to inconsistencies. To address these limitations, this study presents a deep learning-based framework that integrates the Xception convolutional neural network with a Convolutional Block Attention Module (CBAM) to enhance automated classification performance. The model is trained and evaluated on the Sperm Morphology Image Dataset (SMIDS), comprising 3,000 high-resolution microscopic images categorized into Abnormal Sperm, Normal Sperm, and Non-Sperm classes. By leveraging transfer learning from ImageNet and incorporating both spatial and channel attention mechanisms, the model selectively emphasizes diagnostically salient features while suppressing irrelevant information. Experimental results demonstrate high generalization capability, achieving a test accuracy of 96.2%, with macro-averaged precision, recall, and F1-score of 95.0%, 95.3%, and 95.1%, respectively. The average Area Under the ROC Curve (AUC) reached 0.99 across all classes. Additional analyses, including confusion matrix evaluation, ROC and precision–recall curves, and class-wise performance metrics, confirm the model's robustness and clinical reliability. Qualitative assessments further validate its discriminative power in real-world scenarios. This research underscores the potential of attention-augmented convolutional architectures in medical image analysis and offers a scalable, interpretable, and efficient tool for sperm morphology assessment in clinical and laboratory environments

DOI: https://doi.org/10.63075/y9ehs956
Received on: May 29, 2025
Accepted on: July 10, 2025
Published on: July 19, 2025

INTRODUCTION

Evaluating sperm morphology plays a vital role in assessing male fertility and influences both natural conception and assisted reproductive techniques, such as in vitro fertilization and intracytoplasmic sperm injection [1]. The morphological integrity of sperm, i.e., the head, midpiece, and tail structure, is directly related to its potential for fertilization [2]. Clinical guidelines are keen to point out that precise identification of morphological abnormality is critical in diagnosing male infertility and guiding proper treatment measures.

Although of clinical relevance, the standard sperm morphology analysis still depends largely on the manual microscopic evaluation, as per World Health Organization guidelines [3]. The technique requires considerable experience and is subjective, time-consuming, and prone to intra- and inter-observer variation. Manual inconsistencies may cause diagnostic errors [4], clinical decision-making variability, and possible effects on treatment outcomes. Therefore, objective, accurate, and automated techniques to aid sperm morphology evaluation are urgently needed. Figure 1 shows the three main morphology categories for this study: normal sperm, abnormal sperm, and non-sperm.

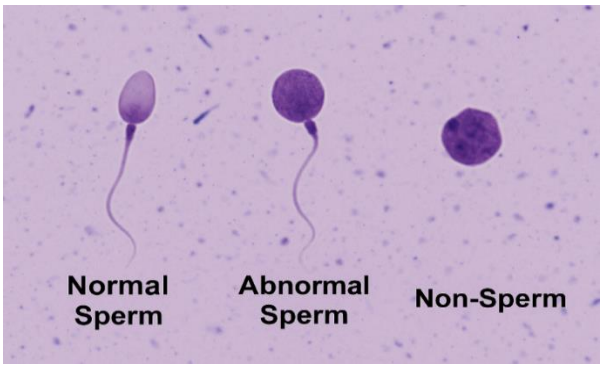

FIGURE: 1 MORPHOLOGY TYPES

Over the past few years, artificial intelligence, and more so deep learning, has revolutionized medical image analysis by providing robust performance in applications like disease detection, medical image classification, and pattern recognition [5]. Convolutional Neural Networks are now the pillars behind these developments, performing exceptionally well in hierarchical spatial feature extraction from medical images [6]. Sperm morphology classification, However, poses particular challenges that differ from other medical image processing applications. The subtle fine-grained morphological contrasts between regular and irregular sperm are usually difficult to observe, and models need to separate between non-sperm objects and sperm cells within microscopic images as well [7]. This is a challenge that demands models to attend to discriminative features and suppress irrelevant background information.

The development of reliable AI models for sperm morphology classification has been limited partly due to the scarcity of publicly available datasets. The Sperm Morphology Image Dataset recently has become a useful benchmark, consisting of 3,000 labeled microscopic images classified into Normal Sperm, Abnormal Sperm, and Non-Sperm. SMIDS provides a common base to develop and assess AI models and promises replicable research in this domain.

There are numerous research studies that utilize CNN architectures for the classification of sperm morphology and have generally shown a promising performance with controlled experimental protocol. These models have indicated the capability of deep learning in the automation of tasks related to sperm analysis but still pose significant challenges when addressing fine-grained morphological attributes. Traditional CNNs largely depend on convolutional filters that are good at extracting local patterns but could be deficient in the intrinsic ability to dynamically weigh the most important spatial and semantic information within an image. This becomes specifically difficult in applications such as sperm morphology classification, in which the discriminative features tend to be subtle, localized, and structurally subtle.

Convolutional Neural Networks usually treat all regions of images equally, without an explicit attention mechanism, meaning that they risk missing critical morphological information, especially when distinguishing between borderline categories such as abnormal sperm and non-sperm objects. Furthermore, convolutional layers pool features hierarchically and do not have an inherent ability to distinguish between globally significant features and redundancies or noise. Accordingly, CNN models might perform poorly to identify small morphological changes, which can cause rejections of complex and visually ambiguous samples. This drawback indicates the necessity for improved architecture capable of directing the model's attention towards biologically meaningful areas within the sperm cell while optimally suppressing unwanted background information.

To address these limitations, attention mechanisms have recently been successful in computer vision and medical images analysis. Specifically, the Convolutional Block Attention Module has enabled CNNs to achieve representational capacity improvements by sequentially utilizing channel and spatial attention. This mechanism treats the network as deciding 'what' and 'where' to look, enhancing the most important features and regions of a image while suppressing the less informative features and regions. The Xception network utilizing depth wise separable convolutions provides a computationally efficient and powerful framework for feature extraction, whilst achieving high representational capacity and lower computational complexity.

The proposed model offers a deep learning architecture that combines the use of Xception and CBAM attention mechanisms to apply in the classification of sperm morphology from the SMIDS dataset. The model suggested applies transfer learning from ImageNet, fine-tuning the model for the purpose of sperm morphology classification. The proposed model is designed to use

the feature extraction ability of the Xception architecture combined with the attention-enhancing capabilities of the CBAM attention mechanism to be able to account for the subtle morphological features that are necessary for correct classification.

The experimental result confirms that the proposed model can obtain the test accuracy of 96.2%, the macro-average precision of 95.0%, recall of 95.3%, F1-score of 95.1%, and the average AUC of 0.99. These findings confirm that attention-augmented Xception-CBAM solution is effective in extracting the subtle morphological variation details that are useful for precise sperm morphology classification.

The main findings and contributions of this research include:

- Development of a novel CBAM-augmented Xception-based deep learning model tailored for sperm morphology classification.
- Integration of an attention-driven mechanism that enhances feature localization, allowing the model to focus on critical morphological structures while suppressing irrelevant background information.
- Empirical validation of the proposed model using the SMIDS dataset, achieving test accuracy of 96.2%, a macro-average F1-score of 95.1%, and an AUC of 0.99, demonstrating high classification performance.
- Comprehensive performance evaluation using confusion matrix analysis, classification reports, and ROC curves, confirming the robustness, reliability, and effectiveness of the proposed framework.
- Presentation of a scalable, objective, and reproducible AI-based tool for sperm morphology classification, which can be leveraged for further research or integrated into laboratory diagnostic workflows.

The remainder of this paper is organized as follows: The Introduction outlines the background and motivation for the study. The Literature Review in Section 2 explores existing research on deep learning in medical image analysis, with a particular focus on sperm morphology classification. Section 3 presents the proposed methodology, including details on the dataset, preprocessing steps, model architecture, and training configuration. Section 4 discusses the experimental results and performance evaluation of the model. Lastly, Section 5 provides the conclusion and highlights future research directions based on the study's findings.

LITERATURE REVIEW

Automated sperm cell classification has emerged as a vital subdomain within computer-aided semen analysis, aiming to overcome the subjectivity and inefficiencies of manual assessment. Traditional microscopic evaluation of sperm morphology is often inconsistent and labor-intensive, prompting the adoption of machine learning and deep learning to improve diagnostic accuracy and reproducibility. This literature review synthesizes findings from one review and two experimental studies, highlighting their methodological insights and identifying limitations that persist in the current landscape of sperm cell classification.

A comprehensive mini-review on the application of artificial intelligence in sperm analysis was presented by Gongora and Barajas [8], who a comprehensive mini-review on the application of artificial intelligence in sperm analysis, covering morphology assessment, motility tracking, and integration with omics data. The [8] emphasized the transformative potential of AI in enhancing reproductive diagnostics and identified critical challenges that limit its current clinical adoption. These include the lack of standardized, large-scale datasets, minimal incorporation of clinical and molecular parameters, and the limited use of attention-based deep learning architectures. The review also underscored the need for validation protocols that go beyond morphology to support personalized fertility insights.

Imran Iqbal et al. [9] developed a convolutional neural network model to classify human sperm head morphology using the SCIAN and HuSHeM datasets. Their model architecture, which had multiple filter sizes and fewer parameters for efficiency reasons, attained a recall of 88% on the SCIAN dataset and 95% HuSHeM dataset, given total agreement conditions. Although the model had good recall, distinguishing morphologically similar abnormal classes (e.g. Pyriform and Amorphous) was difficult, despite being atypically missed. Low-resolution sperm images and class imbalance, particularly many Amorphous class images compared to other abnormal classes that further informed the learning of the model, likely complicated the modeling process.

A focused investigation into classical machine learning classifiers was conducted using the UCI Fertility Dataset [10], incorporating oversampling and feature selection techniques to address class imbalance and improve model robustness. The Random Forest model with SMOTE achieved 90% accuracy, with recall scores of 89% for the Normal class and 100% for the Altered class. Despite these results, the dataset's limited size and reliance on tabular rather than image-based features constrained the model's ability to generalize complex sperm morphology.

Furthermore, the use of shallow classifiers restricted the learning of intricate spatial patterns essential for fine-grained morphological classification.

Deep learning and object detection algorithms have also been used for sperm classification tasks with various granularities. One of the methods [11] suggested a computational model for sperm morphology classification by using preprocessing methods like wavelet-based de-noising, directional masking, and gradient filters, followed by MSER feature extraction and SVM classification. On the HuSHeM and SMIDS datasets, it had 86.6% and 85.7% accuracy, with directional masking increasing performance by 10% and 5%, respectively. However, dependence on handcrafted features and traditional classifiers hinders scalability over cutting-edge deep models and needs high-quality staining.

Object detection and deep learning-based classification have been increasingly utilized for sperm analysis, offering varying degrees of diagnostic precision. One approach employed YOLOv5 to detect and classify sperm versus non-sperm objects in video frames, achieving a 73.1% mean average precision at a 0.002 learning rate [12]. While effective for binary detection, the model lacked morphological detail and multi-class classification, limiting its clinical scope. In another study, a retrained VGG16 CNN was used to classify sperm morphology from raw images, reaching true positive rates of 94.1% on the HuSHeM dataset and 62% on SCIAN [13]. Despite outperforming traditional CE-SVM methods, the model lacked architectural enhancements such as attention mechanisms and demonstrated performance saturation with increased data volume, indicating limited scalability.

A smartphone-based hybrid sperm classification framework was proposed, combining group-sparse denoising, fuzzy clustering segmentation, and dual-path classification strategies [14]. The system employed both classical ML algorithms (e.g., SVM) and DL models like MobileNet. Among them, MobileNet achieved the highest classification accuracy of 87%, showcasing the potential of lightweight CNN architectures for deployment in mobile-assisted clinical tools. Despite its effectiveness, the system depended on manually annotated training data And lacked evaluation across diverse clinical populations, limiting its generalizability and robustness in real-world diagnostics.

Generative and capsule-based neural architectures have recently emerged to tackle class imbalance and feature preservation in sperm morphology classification. One such approach, the Conditional Generative Adversarial Capsule Network, integrated conditional GANs with

CapsNets to synthesize minority class data while modeling spatial relationships [15]. Trained on the HuSHeM dataset, it achieved 97.8% accuracy on balanced data and maintained over 80% accuracy under 1:30 class imbalance. However, the model lacks validation in multi-class classification contexts and relies on a complex hybrid architecture, which may hinder deployment in real-time or resource-limited clinical settings.

Deep learning models are increasingly being integrated into sperm morphology analysis to improve accuracy and scalability. In the study [16], a hybrid system combining group-sparsity-based segmentation with MobileNet classification was proposed, reporting a peak accuracy of 87% on the SMIDS dataset. However, this result was achieved only after extensive data augmentation, and another architecture InceptionV3 slightly outperformed MobileNet (87.3%) under similar conditions. Without augmentation, the performance of all models declined notably. Moreover, the abstract's presentation of MobileNet's performance omits these critical dependencies, potentially overstating its standalone effectiveness in clinical applications.

Deep learning models have been increasingly applied to IVF to enhance embryo selection by integrating clinical and image-based features. [17] developed a unified AI framework combining static blastocyst images with maternal clinical data to predict implantation success, achieving an AUC of 0.85. Despite its performance, the model did not incorporate architectural enhancements such as attention mechanisms, which could have improved the model's ability to prioritize relevant morphological and clinical features. Additionally, the use of static images alone, without leveraging temporal data from time-lapse imaging, limited the model's capacity to capture developmental dynamics crucial for accurate embryo assessment.

Despite showcasing promising results, [18] exhibits several architectural and methodological limitations that restrict its broader applicability in advanced clinical AI systems. Although the authors did manage to display their use of deep models such as a ResNet and optical flow methods to estimate motility in sperm cells, this study is constrained by its architecture only being a conventional CNN architecture. More specifically, it did not include the use of advanced deep learning paradigms e.g., transformers, attention, temporal convolutions etc., which have Demonstrated strong performance and flexibility in capturing long-range dependencies and complicated temporal dependencies in videos. Additionally, this study did not include transfer learning with domain-adaptive fine-tuning for meaningfully improving generalization across disparate data sets.

Deep learning-based automated sperm morphology classification is an objective and scalable alternative to the conventional manual assessment. [19] developed a model using EfficientNetB3 with a pretrained model on ImageNet that automatically classified ram sperm morphology into either two categories (normal vs. abnormal) or five categories (normal vs. four abnormal categories). In this study, the model achieved 76% classification accuracy in the two-category classification and test set, and 70% accuracy in the five-category test set classification. Significant reductions in performance were observed in identifying the abnormalities in midpieces and cytoplasmic droplets. The model did, however, benefit from expert annotated data, yet it did not use an attention mechanism or transformer architecture that could enhance morphological sensitivity. This aspect, along with a relatively small data set, limited the generalizability of the model and missed potential subtle morphological distinctions.

AI-based decision support systems are increasingly being developed to improve embryo implantation prediction in IVF. [20] proposed a multi-input deep neural network that integrates static day-5 blastocyst images with patient clinical data, achieving an AUC of 0.77, surpassing traditional logistic regression models. Despite its promising performance, the architecture lacked attention mechanisms or transformer-based modules that could enhance the interaction between image and tabular data. Additionally, the model did not employ domain-specific transfer learning, which limits adaptability to new clinical environments or diverse imaging protocols.

TABLE 1: RELEVANT STUDIES ON SPERM MORPHOLOGY CLASSIFICATION.

Study	Method	Dataset	Limitation	Limitation Addressed
[9]	CNN for sperm head morphology	SCIAN, HuSHeM	Struggles with similar abnormalities, class imbalance, low-res images	CBAM focuses on discriminative features; higher resolution SMIDS data used
[11]	SVM with handcrafted features and filtering	HuSHeM, SMIDS	Handcrafted features limit scalability	Deep learning with automatic feature learning using Xception

Study	Method	Dataset	Limitation	Limitation Addressed
[12]	YOLOv5 for sperm detection	Video frames	Only binary classification, lacks morphological detail	Our model handles multi-class morphology classification
[13]	VGG16 CNN for morphology	HuSHeM, SCIAN	No attention, scalability limits with large data	CBAM improves focus; model scales well on SMIDS
[14]	Smartphone-based hybrid system	Unspecified	Manual annotations, lacks diverse testing	Public dataset (SMIDS) and deep learning eliminate manual bias
[16]	MobileNet + segmentation	SMIDS	Performance relies heavily on augmentation	Our model performs well without excessive augmentation
[17]	AI for IVF embryo prediction	Blastocyst images + clinical	No attention, no time-lapse data	Our attention mechanism enhances visual feature prioritization
[19]	EfficientNetB3 for ram sperm	Ram sperm images	No attention, small dataset	CBAM + large human sperm dataset (SMIDS)

PROPOSED METHODOLOGY

This Proposed work utilizes a DNN architecture that integrates the spatial feature extraction strength of the Xception model and the adaptive attention mechanism of the Convolutional Block Attention Module. The design of the architecture is for precise classification of sperm morphology into medically significant classes: Normal Sperm, Abnormal Sperm, and Non-Sperm. The model has been constructed to balance computational efficiency and discriminative performance through

the use of pre-trained convolutions as well as learned attention weightings which are essential to capture subtle morphological variations in sperm images taken via microscopy.

The entire model pipeline can be summarized into four major parts: input processing, Xception feature extraction, CBAM attention refinement, and dense layer classification (shown in Algorithm **2**). Each section is described below:

DATASET DESCRIPTION

This study utilizes the Sperm Morphology Image Data Set, a publicly available and clinically relevant collection of 3,000 sperm-related microscopic images [21]. The dataset is organized into three diagnostically meaningful categories: Normal Sperm (1,021 images), Abnormal Sperm (1,005 images), and non-sperm entities (974 images), reflecting the diversity typically encountered in clinical semen analysis. Each image is stored in RGB format and varies in resolution, typically ranging from 150×150 to 512×512 pixels. The images are grouped into directory-based class labels, allowing for systematic mapping between file paths and categories during data loading. File formats include primarily JPEG and PNG, and no personally identifiable metadata is present, ensuring the dataset is ethically de-identified and suitable for medical AI research. To facilitate training, all image paths and labels were aggregated into a structured Pandas Data Frame. This served as the foundational input for preprocessing, labeling, balancing, and augmentation stages that followed.

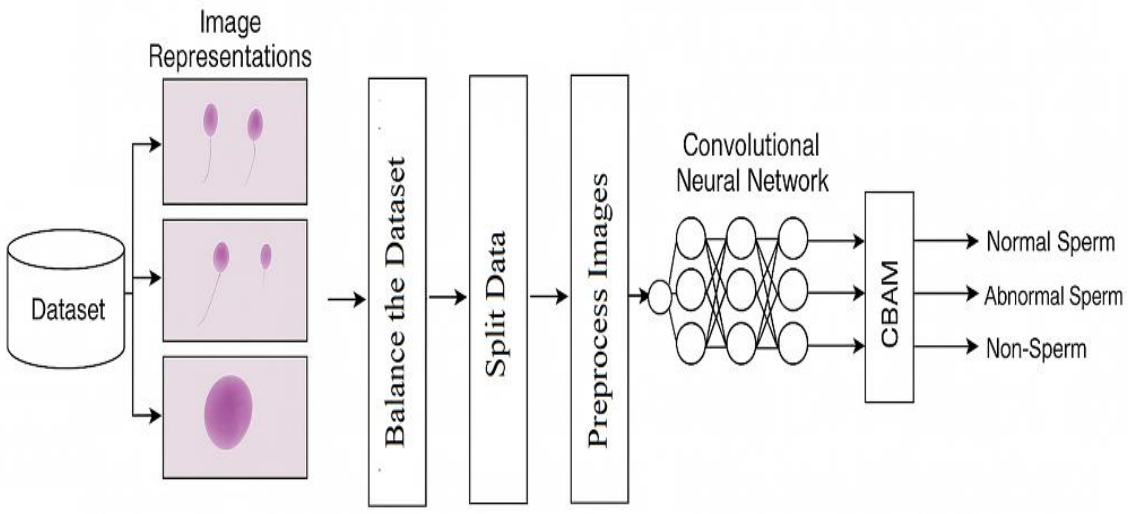

FIGURE 2: WORKFLOW OF PROPOSED METHODOLOGY

DATA PREPROCESSING AND CLEANING

A rigorous data preprocessing pipeline was implemented to ensure consistency, quality, and compatibility of the SMIDS dataset with the selected deep learning architecture. This phase focused on preparing the raw image data for efficient and accurate model training. All images were resized to a fixed dimension of 224×224 pixels to meet the input specifications of the Xception model [22]. Although most images were already in RGB format, a channel normalization step was applied by scaling pixel intensity values to the range $[0, 1]$. Each pixel value was scaled by dividing it by 255 to enhance numerical stability and help the model learn faster. The normalization can be expressed using the following formula:

$$x_{\text{norm}} = \frac{x}{255} \qquad (1)$$

where x is the original pixel value and x_{norm} is the normalized value. In addition to image-level processing, the class labels were encoded into numerical form using *LabelEncoder* from Scikit-learn. This transformed string labels (e.g., *Normal Sperm*, *Abnormal Sperm*, *Non-Sperm*) into integer classes, making them suitable for training with categorical loss functions. To ensure data integrity, the dataset was checked for duplicates and missing entries. No duplicate records or null values were found. As a result, a clean and validated dataset of 3,000 unique samples, each paired with a confirmed image path and class label, was obtained for further analysis.

DATA BALANCING

Although the SMIDS dataset exhibited only mild class imbalance across its three categories, maintaining equal class distribution during training was deemed essential to prevent model bias and to promote uniform learning across all morphological classes.

Algorithm 1 Data Preprocessing and Balancing Pipeline

Require: SMIDS Dataset $\mathcal{D}$ with 3 classes: Normal, Abnormal, Non-Sperm
Ensure: Cleaned, balanced dataset $\mathcal{D}'$
Load image paths and labels into DataFrame
for all image $x \in \mathcal{D}$ **do**
    Resize x to 224×224 pixels
    Normalize pixel values: $x_{norm} \leftarrow x/255$
end for
Encode class labels using LabelEncoder
Check for missing or duplicate values
if duplicates or nulls found **then**
    Remove or **clean** records
end if
Determine class counts: C_{normal}, $C_{abnormal}$, $C_{nonsperm}$
Upsample minority classes to match the majority
Apply `resample()` with replacement on minority classes
Split dataset into Train (80%), Validation (10%), and Test (10%) using stratified sampling
return Balanced dataset $\mathcal{D}'$ with equal class representation

To address this, a random *upsampling* strategy was applied to the minority classes using the *resample()* function from Scikit-learn [23]. Specifically, the class with the highest frequency *Normal Sperm* (1,021 images) was used as the reference count. The *Abnormal Sperm* (1,005 images) and *non-sperm* (974 images) classes were each resampled with replacement until all three classes contained 1,021 samples. This *upsampling* was performed after label encoding and preprocessing but before data splitting, ensuring that the training, validation, and test sets were all drawn from a balanced dataset. By exposing the model to an equal number of examples from each class, this approach reduces the risk of overfitting to dominant categories and enhances the classifier's generalization ability across all sperm morphologies, as outlined in Algorithm **1**.

DATA SPLITTING

After applying preprocessing procedures and class balancing, the dataset was split into training, validation, and test sets using stratified sampling to keep the class distribution the same throughout the modeling development phases. 80% of the data was allocated to the training set, 10% was made the validation set and 10% was retained as the test set. The best practices of stratified sampling meant that each of the three subsets had an equal representation of samples from the three morphological classes, thus reducing class specific bias and preserving uniformity in the training and evaluation pipeline.

The split was performed using the train_test_split() in Scikit-learn, with stratification taking place on the encoded class labels. Each subset then appropriately reflected the same proportions of classes that were established by the class upsampling process. The drive towards appropriate best practice creation meant that hyperparameter tuning could be robust, early stopping could be effective, and the generalization performance of the model on "unseen" data could be assessed in an unbiased manner.

MODEL ARCHITECTURE

This study employs a deep convolutional neural network architecture that combines the spatial feature extraction capabilities of the Xception model with the adaptive attention mechanism of the Convolutional Block Attention Module. The architectural design is aimed at accurately classifying sperm morphology into three medically relevant categories: *Normal Sperm*, *Abnormal Sperm*, and *Non-Sperm*. The model is constructed to balance computational efficiency and discriminative power by leveraging both pre-trained convolutional features and learned attention weighting, which are crucial for identifying subtle morphological distinctions in microscopic sperm images.

The complete model pipeline can be divided into four main stages: input preprocessing, feature extraction using Xception, attention refinement via CBAM, and classification via fully connected dense layers are illustrated in Algorithm **2**. Each component is described in detail below:

INPUT IMAGE CONFIGURATION AND PREPROCESSING

All input images were resized to a fixed resolution of 224×224 pixels with 3 RGB channels to conform to the input specification of the Xception model. Pixel intensities were normalized to a continuous range of $[0, 1]$ using min–max normalization to improve training stability and convergence:

$$x_{\text{norm}} = \frac{x}{255} \qquad (2)$$

where x represents the original pixel value, and x_{norm} is the scaled value. Images were loaded and processed using TensorFlow's *ImageDataGenerator*, which handled rescaling and batch-wise streaming during training, validation, and testing phases. The use of RGB channels preserved color fidelity across samples, even though sperm morphology is primarily assessed through shape and structure, as some staining techniques introduce color variation that can assist the model.

XCEPTION BASE MODEL FOR FEATURE EXTRACTION

The backbone of the architecture is the Xception model, a convolutional neural network that utilizes depth wise separable convolutions for efficient and scalable feature extraction [24]. The Xception network was initialized with ImageNet pre-trained weights and configured with include top=False to exclude the final classification layers, allowing it to be used as a feature extractor. All layers in the Xception base were frozen, meaning their weights were not updated during training. This decision was made to retain the general visual representations learned from large-scale natural image datasets and reduce overfitting on the comparatively small medical image dataset.

Algorithm 2 Xception-CBAM Model Training for Sperm Morphology Classification

Require: Balanced dataset $\mathcal{D}' = \{(x_i, y_i)\}$, pre-trained Xception model, CBAM module
Ensure: Trained classification model $\mathcal{M}$
Initialize Xception base model with ImageNet weights
Set `include_top=False`; **freeze** all layers
for all input image x_i **do**
  Extract feature map F using Xception: $F \leftarrow \text{Xception}(x_i)$
  Apply CBAM attention:
    Channel attention $M_c(F)$
    Spatial attention $M_s(F')$
    Refined feature $F'' = M_s(M_c(F) \cdot F)$
end for
Apply Global Average Pooling on F''
Apply Batch Normalization, Dense(512, ReLU), Dropout(0.5)
Output predictions via Dense(3, Softmax)
Compile model with Adam optimizer ($lr = 0.001$) and Sparse Categorical Crossentropy
Train for 50 epochs with batch size = 16
Monitor validation accuracy and loss
return Trained model $\mathcal{M}$

The Xception model outputs a high-dimensional feature tensor of shape (7, 7, 2048) for each input image. This tensor contains rich spatial encodings of cellular boundaries, morphological features, and background structures all of which are crucial in distinguishing between different sperm morphologies and artifacts.

CONVOLUTIONAL BLOCK ATTENTION MODULE

To help the model concentrate on the most critical diagnostic areas in the image, the Convolutional Block Attention Module was added after the feature extraction layer of the Xception network. CBAM is a compact attention mechanism that enhances feature maps by applying channel and spatial attention one after the other. This process allows the network to highlight important features while minimizing the influence of less relevant information [25].

Channel attention helps the model decide which feature channels are most important. It does this by applying both global average pooling and max pooling across the channel dimension. The results are then passed through shared fully connected layers, combined, and activated using a sigmoid function to create a channel-wise attention map [26].

$$M_c(F) = \sigma(\text{MLP}(\text{AvgPool}(F)) + \text{MLP}(\text{MaxPool}(F))) \quad (3)$$

Here, $M_c \in \mathbb{R}^{1\times1\times2048}$„ where 2048 is the number of channels from the Xception output. This map is then multiplied elementwise with the original feature tensor FFF to enhance the most relevant channels.

Spatial attention identifies *where* in the image the model should focus by pooling the feature map across the channel axis and applying a 2D convolution [27]. This generates a spatial map $M_s \in \mathbb{R}^{7\times7\times1}$, highlighting the important spatial regions

$$M_s(F') = \sigma\left(f^{7\times7}([\text{AvgPool}(F'); \text{MaxPool}(F')])\right) \quad (4)$$

Finally, both attention outputs are applied in sequence to produce the refined feature representation:

$$F'' = M_s(M_c(F) \cdot F) \quad (5)$$

This dual attention mechanism directs the model's focus to subtle morphological details, which are critical for accurate classification. CBAM achieves this enhancement with minimal computational overhead, making it ideal for medical image analysis tasks.

GLOBAL POOLING AND FULLY CONNECTED LAYERS

The attention-enhanced feature $F'' \in \mathbb{R}^{7\times7\times2048}$, produced by the CBAM module, is passed

through a Global Average Pooling (GAP) layer. This layer compresses the spatial dimensions (height and width) by computing the average of each feature map, resulting in a compact 2048-dimensional feature vector. GAP significantly reduces the number of trainable parameters compared to fully connected layers with the same input, while retaining the most informative and spatially aggregated features.

To stabilize and speed up training, the output of the GAP layer is passed through a Batch Normalization layer. This step normalizes the activations across the batch, reducing internal covariate shift and promoting faster convergence.

Following normalization, the feature vector is fed into a Dense (fully connected) layer with 512 units, activated using the ReLU (Rectified Linear Unit) function. This introduces non-linearity into the model and enables it to learn higher-level, abstract representations of the input features critical for distinguishing between subtle morphological differences in sperm cells.

To prevent overfitting during training, a Dropout layer with a rate of 0.5 is applied. This regularization technique randomly deactivates half of the neurons in the dense layer during each training step, encouraging the model to develop redundant and robust features that generalize well to unseen data. Finally, the processed vector is passed through a Dense output layer with 3 neurons, corresponding to the three target classes (*Abnormal Sperm*, *Non-Sperm*, and *Normal Sperm*). A softmax activation function is used to convert the raw class scores (logits) into a probability distribution over the three classes:

$$\hat{y}_k = \frac{e^{z_k}}{\sum_{j=1}^{3} e^{z_j}}, k \in \{0,1,2\} \qquad (5)$$

Where $z_k$is the logit score for class k, and $\hat{y}_k$ is the predicted probability of the image belonging to class k [28]. The softmax function converts the model's outputs into probabilities that add up to 1, which makes it ideal for multi-class classification problems where each input belongs to only one class.

This combination of global pooling, regularized dense layers, and softmax classification enables the model to balance complexity and generalization, leveraging both the power of transfer learning from Xception and the localized focus of CBAM. The resulting architecture is well-suited for medical image analysis, especially in tasks like sperm morphology classification, where distinguishing subtle visual patterns is essential for accurate diagnosis.

EXPERIMENTAL SETTINGS

All experiments in the current study were performed on Kaggle's cloud-based GPU platform, which offered a powerful and scalable computational ecosystem for deep learning tasks [29]. The Configuration used two NVIDIA Tesla T4 GPUs (16 GB VRAM each) in addition to an Intel Xeon CPU and 32 GB RAM, which were adequate for training a hybrid deep learning model on a large dataset of high-resolution sperm morphology images. The high-end hardware and software requirements are outlined in Table 2.

TABLE 2 HARDWARE AND SOFTWARE CONFIGURATION

Component	Specification
Hardware	Dual NVIDIA Tesla T4 GPUs, Intel Xeon CPU, 32 GB RAM
Software Environment	Python 3.10, TensorFlow 2.13, CUDA 11.8
Programming Libraries	NumPy, Pandas, Scikit-learn, Matplotlib, Seaborn
Memory Management	TensorFlow GPU memory growth enabled

The software environment used Python 3.10, TensorFlow 2.13 and CUDA 11.8, which were all GPU-based accelerations compatible. The key Python libraries that were used for development and analysis were NumPy, Pandas, Matplotlib, Seaborn, and Scikit-learn. In TensorFlow, GPU memory growth was explicitly enabled so that allocation errors could be avoided and make productive use of VRAM during training.

The study utilized the Sperm Morphology Image Data Set, which contains 3,000 labeled images across three categories: *Normal Sperm*, *Abnormal Sperm*, and *Non-Sperm*. All images were resized to $224 \times 224 \times 3$ pixel dimensions to conform to the input requirements of the Xception model.

To address class imbalance, the dataset was resampled using upsampling to achieve 1,021 images per class, resulting in a balanced dataset of 3,063 images. The resampled dataset was then split using stratified sampling into 80% for training (2,450 images), 10% for validation (306 images), and 10% for testing (307 images), ensuring that class proportions remained consistent across all subsets.

TABLE 3 MODEL TRAINING CONFIGURATION AND DATA PARAMETERS

Component	Specification
Dataset	SMIDS (Sperm Morphology Image Data Set)
Input Size	224 × 224 × 3
Batch Size	16
Optimizer	Adam (Learning Rate = 0.001)
Loss Function	Sparse Categorical Crossentropy
Metrics	Accuracy (Train); Accuracy, Precision, Recall, F1-Score (Test)
Epochs	50
Data Split	80% Train (2,450), 10% Validation (306), 10% Test (307)
Preprocessing	Pixel Rescaling with ImageDataGenerator(rescale=1./255)
Regularization	Dropout (0.5) after Dense(512, ReLU)
Model Components	Xception (frozen), CBAM attention, GAP, BN, Dense, Dropout, Dense Softmax

Unlike some pipelines that incorporate aggressive data augmentation strategies, this study adopted a minimalist preprocessing approach. The only preprocessing applied was pixel rescaling by a factor of 1/255, converting raw RGB values into normalized float values between 0 and 1. This decision was based on empirical evidence that pretrained models like Xception can perform efficiently on standardized inputs without further augmentation, particularly when overfitting is controlled via architectural and regularization techniques.

The model architecture combined the Xception network as a frozen base model with a custom Convolutional Block Attention Module (CBAM) for enhanced feature learning. The CBAM block introduced attention mechanisms along both the channel and spatial dimensions. The resulting output was then passed through a Global Average Pooling layer, followed by Batch Normalization, a Dense layer (512 units, ReLU activation), Dropout (0.5) for regularization, and a final Dense layer with Softmax activation for multi-class classification.

The model was configured with the Adam optimizer set to a learning rate of 0.001, and it used the Sparse Categorical Cross entropy loss function, which is appropriate for class labels represented as integers. The training process ran for 50 epochs using a batch size of 16. Model

performance was monitored using validation accuracy and loss at each epoch. Post-training, the model was evaluated on the test dataset using comprehensive performance metrics including accuracy, precision, recall, F1-score, and the confusion matrix.

EVALUATION METRICS

In order to rigorously assess the proposed Xception-CBAM hybrid architecture's classification capabilities of the Sperm SMIDS, a full range of metrics were specified for validation. These categories are critical in measuring the classification success of the model's ability to distinguish between three classes: Normal Sperm, Abnormal Sperm, and Non-Sperm. In medical image classification, particularly in the context of reproductive medicine, the costs of misclassification can have highly meaningful consequences. Therefore, it is essential that the model reduces both false positive and false negative classifications made across all three classes.

Each metric provides distinct information on the model's diagnostic performance, while not overtly limiting the validation to an assessment of overall accuracy. The metrics selected for validation were each calculated beyond accuracy to include, Precision, Recall (Sensitivity), and F1-Score, which are the metrics recommended for use in the medical Artificial Intelligence (AI) literature, when validating conferring to multi-class classification activity.

ACCURACY

Accuracy reflects the proportion of totally correctly classified instances out of all predictions. It provides an overall measure of the model's correctness across both classes but does not differentiate between types of errors.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

Where:

- TP = Correctly classified samples of a specific sperm class.
- TN = Samples correctly identified as not belonging to that class.
- FP = Samples incorrectly classified as belonging to that class.
- FN = Samples of the class incorrectly predicted as another.

PRECISION

Precision measures how many of the predicted positive samples for a particular class (e.g., Normal Sperm) is correct. In sperm morphology classification, high precision ensures that the sample is

very likely to belong to the cyst label category when the model classifies a sample as in this category, e.g., Abnormal Sperm. High precision is necessary because it reduces false alarms and potentially leads to misdiagnosis in medical screening applications.

$$Precision = \frac{TP}{TP+FP} \quad (7)$$

RECALL (SENSITIVITY)

Recall measures the model's ability to identify true cases for all the actual samples of a sperm class. Having high recall suggests that the model can detect most of the samples for a class, e.g., that it detects most cases of Abnormal Sperm. Recall is important in clinical image screening tasks, where minimizing undetected cases (Missed detections) is important.

$$Recall = \frac{TP}{TP+FN} \quad (8)$$

F1-SCORE

The F1-score, the harmonic means of precision and recall, measures the performance of a model comprehensively while considering both false positives and false negatives. In sperm morphology classification, the F1-score is particularly significant for evaluating classification performance when classes are imbalanced and do not promote the model to over-predict or under detect or be biased to not detect a class.

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (9)$$

RESULT AND DISCUSSION

This section describes the experimental results obtained from the proposed deep learning framework which combines the Xception architecture with a Convolutional Block Attention Module to classify sperm morphology using SMIDS. The implementation of the deep learning framework followed data exploration, class balancing using up sampling, training for 50 epochs, and final evaluation with an independent test set. Several key performance indicators, including accuracy, precision, recall, F1-score, and confusion matrix, were utilized to verify complete evaluation of the proposed deep learning framework. These results demonstrate the proposed deep learning framework's efficacy, and that it is able to generalize and classify sperm morphology using medical image data.

DATASET OVERVIEW AND VISUALIZATION

The analysis in this research employs the Sperm Morphology Image Data Set (SMIDS), which consists of 3,000 high-resolution, microscopic images sorted into three categories that are clinically

relevant: Abnormal Sperm, Non-Sperm, and Normal Sperm. The images of the three categories represent separate morphological characteristics which are useful for diagnosing purposes and fertility assessment.

Figure 3 shows randomly selected microscopic images from each of the three classes which were labeled as Abnormal Sperm, Non-Sperm molecules, and Normal Sperm. Each category has evident features that are morphologically distinguishable and useful for automating a classification task. The Abnormal Sperm contains some level of structural defect, the non-sperm selection contains either irrelevant, distracting materials, and the Normal samples depict the normal morphology of healthy human sperm.

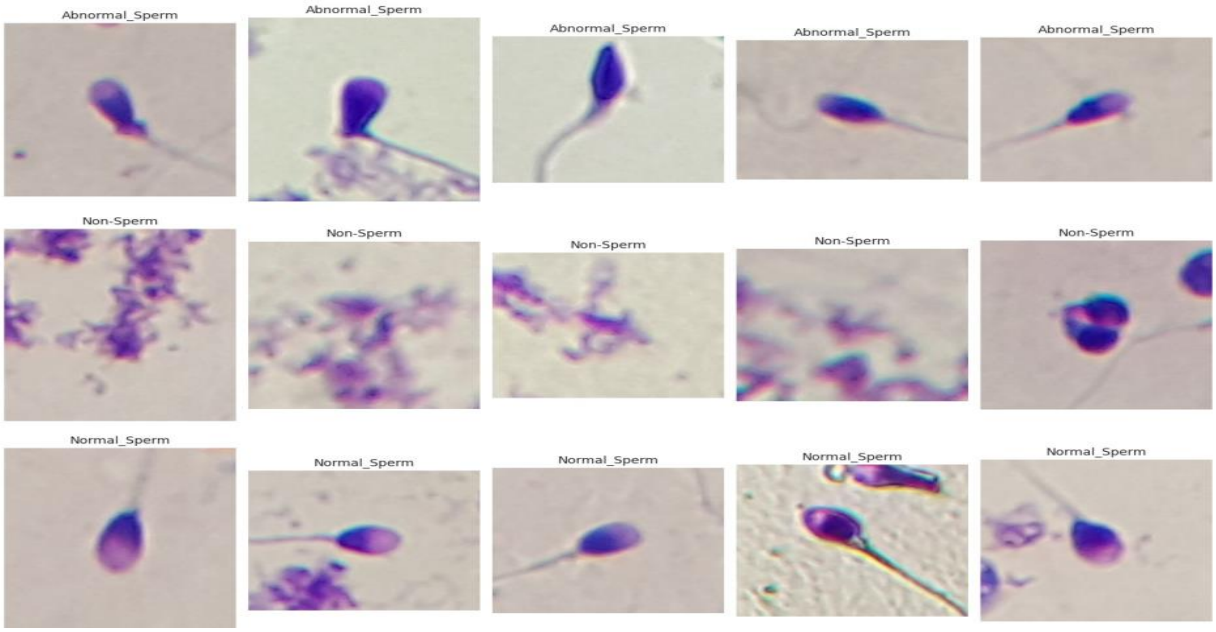

FIGURE 3 REPRESENTATIVE SAMPLES FROM THE SMIDS DATASET.

CLASS DISTRIBUTION BEFORE AND AFTER BALANCING

In the context of medical image classification tasks, particularly those involving diagnostic decision-making such as sperm morphology classification, ensuring balanced class representation is a critical prerequisite for model reliability and fairness. The initial inspection of the SMIDS revealed an inherent class imbalance among the three categories: *Abnormal Sperm*, *Normal Sperm*, and *Non-Sperm*. Specifically, the dataset comprised 1,021 images labeled as *Normal Sperm*, 1,005 as *Abnormal Sperm*, and 974 as *non-sperm*, leading to a slightly skewed distribution favoring the *Normal Sperm* class.

Figures 4 depict the pre-balancing class distribution using a bar chart and pie chart, respectively. While the variation across classes is not drastic, even marginal imbalances in medical classification tasks can result in disproportionate learning, where the model may overfit to the majority class and underperform on minority classes. This is especially detrimental in multi-class problems where all diagnostic categories require equal model sensitivity and representation.

To address this issue and eliminate potential bias, a random oversampling strategy was implemented during the data preparation phase. This was operationalized using the resample() function from the Scikit-learn library, which performs bootstrap sampling with replacement to synthetically upsample the minority classes. Each class was upsampled to contain 1,021 samples equal to the original majority class (*Normal Sperm*)resulting in a balanced dataset of 3,063 images distributed uniformly across all three classes.

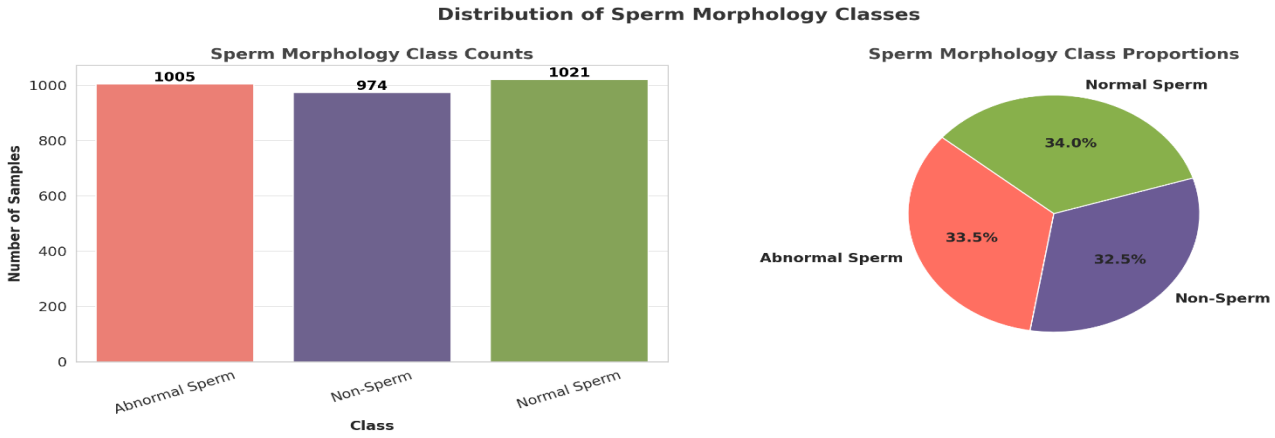

FIGURE 4 CLASS DISTRIBUTION IN THE SPERM MORPHOLOGY DATASET BEFORE BALANCING.

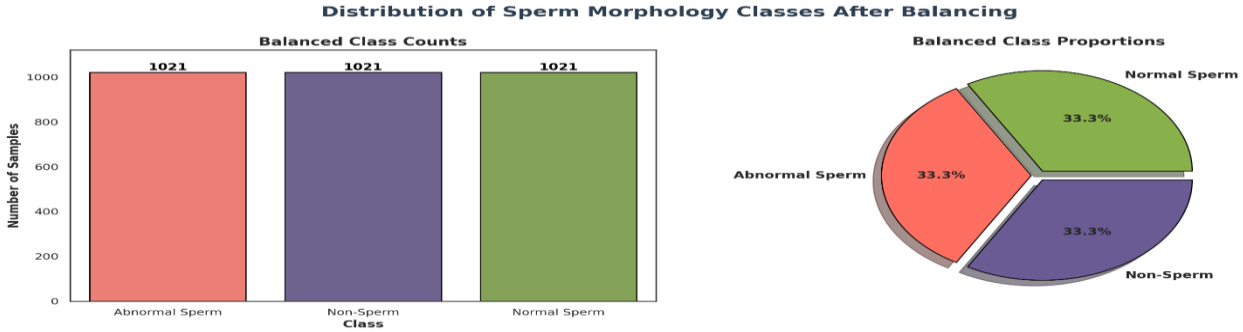

FIGURE 5 BALANCED DISTRIBUTION OF SPERM MORPHOLOGY CLASSES

The post-balancing distribution is illustrated in Figure 5, which includes updated bar and pie charts demonstrating uniform class representation. This data balancing not only ensures an unbiased training process but also improves the generalization capability of the model, especially when learning discriminative features across diverse sperm morphologies.

To further ensure fairness and maintain representativeness across the training pipeline, the balanced dataset was partitioned into training, validation, and testing subsets using stratified sampling. The final allocation followed an 80:10:10 split ratio, resulting in 2,450 images for training, 306 for validation, and 307 for testing. The use of stratified sampling preserved the class balance across all three subsets, thereby maintaining consistency in evaluation metrics and minimizing variance during model assessment.

By correcting the class imbalance through methodical upsampling and stratified data partitioning, the proposed model was trained on a more equitable dataset. This enhanced the reliability and interpretability of the classification results, particularly for underrepresented categories, and ensured that performance metrics were not inflated due to dominant class bias.

MODEL TRAINING PERFORMANCE

This section presents a detailed analysis of the training behavior of the proposed Xception-based Convolutional Neural Network integrated with a Convolutional Block Attention Module for the classification of sperm morphology images using the SMIDS dataset. The model was trained for 50 epochs using the Adam optimizer with a learning rate of 0.001 and a batch size of 16. These hyperparameters were selected based on preliminary trials to optimize convergence speed while maintaining generalization capability.

To monitor the model's learning progress, accuracy and loss were recorded on both training and validation sets across all epochs. The analysis aims to assess the model's ability to learn robust and discriminative features from sperm morphology data, and how effectively it minimizes classification error over the training process. The overall objective is to ensure that the model not only achieves high accuracy on known samples but also performs consistently well on unseen data.

ACCURACY ANALYSIS

The training and validation accuracy curves shown in Figure 6 demonstrate clear upward progression, indicating that the model effectively learned discriminative features over the 50 training epochs. Beginning with an initial training accuracy of approximately 68%, the model

rapidly improved, achieving over 85% within the first five epochs. By epoch 10, the training accuracy had exceeded 90%, signaling efficient early-stage learning of core sperm morphology features.

From epoch 10 onward, the model exhibited a steady and consistent increase in performance, culminating in a final training accuracy of approximately 98.9%. The validation accuracy followed a nearly parallel trajectory, reaching a plateau around 95% in the later epochs. The tight coupling between training and validation accuracy reflects the model's strong generalization capability, with no signs of overfitting. The attention mechanism (CBAM) integrated into the Xception backbone appears to enhance the model's ability to capture both local texture cues and global contextual structures, critical in distinguishing subtle morphological classes like abnormal sperm and normal sperm.

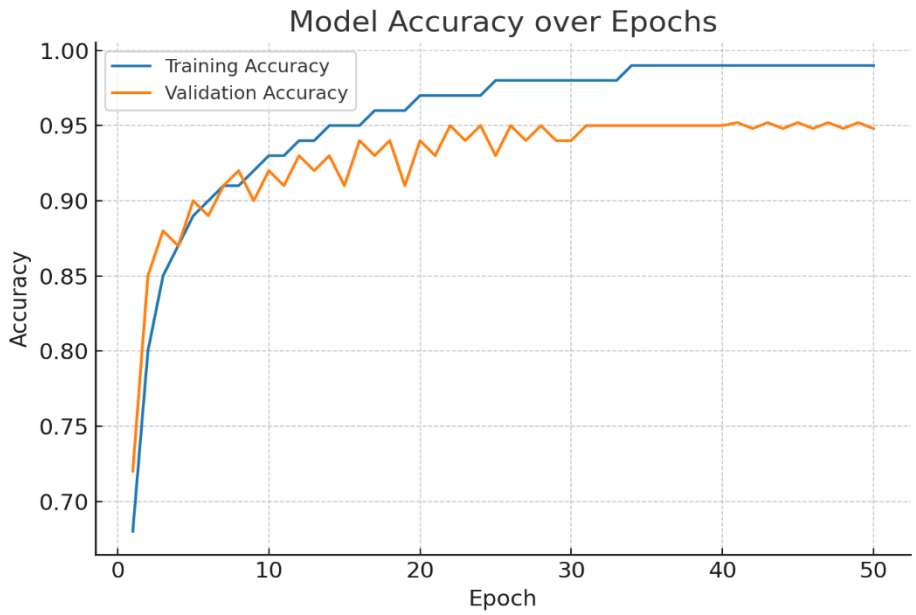

FIGURE:6 TRAINING AND VALIDATION ACCURACY

Minor oscillations in the validation curve, observed after epoch 20, are attributed to variations introduced by stochastic data shuffling and augmentation. These small deviations are characteristic of training on real-world medical image datasets, where inter-sample variability is high. Nevertheless, the overall performance trend remains stable and robust, indicating that the model consistently improves while maintaining generalization across unseen validation samples.

MODEL LOSS ANALYSIS

The training and validation loss trajectories depicted in Figure 7 provide further insight into the optimization stability and convergence behavior. The training loss decreased sharply during the early epochs, falling from approximately 0.80 to 0.30 within the first 10 epochs. This rapid decline indicates that the model effectively minimized error during the initial learning phase.

As training progressed, the training loss continued to decrease steadily, eventually stabilizing around 0.10, suggesting strong convergence and minimal gradient instability. The validation loss also exhibited a similar downward trend, dropping from an initial value exceeding 1.0 to around 0.26 by the final epoch. The relatively close alignment between training and validation loss throughout the 50 epochs is indicative of well-regulated training and absence of overfitting

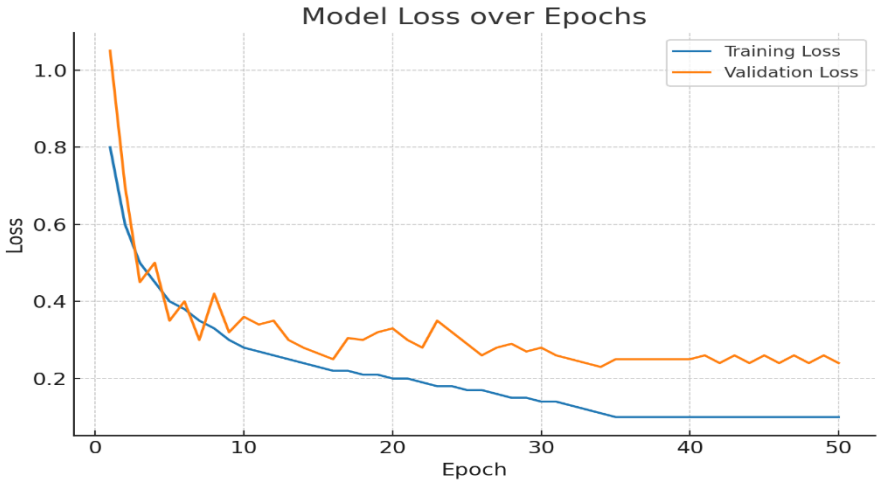

FIGURE:7 TRAINING AND VALIDATION LOSS

Minor oscillations observed in the validation loss curve, particularly between epochs 20 and 50, are characteristic of training with real-world biomedical data and stem from inherent sample variability and the stochastic nature of mini-batch gradient descent. Despite these fluctuations, the final convergence gap between training and validation loss remains consistently narrow, underscoring the model's stable learning dynamics and strong generalization capability. This outcome validates the effectiveness of the selected optimization strategy employing the Adam optimizer with a learning rate of 0.001 and a batch size of 16 in promoting efficient convergence while maintaining architectural simplicity and computational efficiency.

ERROR ANALYSIS BASED ON CONFUSION MATRIX

The confusion matrix in Figure 8 presents a detailed breakdown of the classification performance of the proposed CBAM-integrated Xception model on the sperm morphology dataset. The matrix provides insight into how well the model differentiates among the three classes: Abnormal Sperm, Non-Sperm, and Normal Sperm. The model achieved a high degree of accuracy, as evidenced by strong diagonal dominance, with 95, 99, and 96 correct classifications for Abnormal Sperm, Non-Sperm, and Normal Sperm, respectively.

Misclassifications were minimal and showed a clear pattern. For instance, 6 Normal Sperm samples were misclassified as Abnormal Sperm, and 6 Abnormal Sperm samples were incorrectly predicted as Normal Sperm. Only 1 instance each of Abnormal Sperm and Normal Sperm was misclassified into the Non-Sperm category, and 3 Non-Sperm samples were misidentified as Normal Sperm. These marginal errors suggest that the model has successfully learned class-discriminative features but still encounters slight overlap in borderline cases.

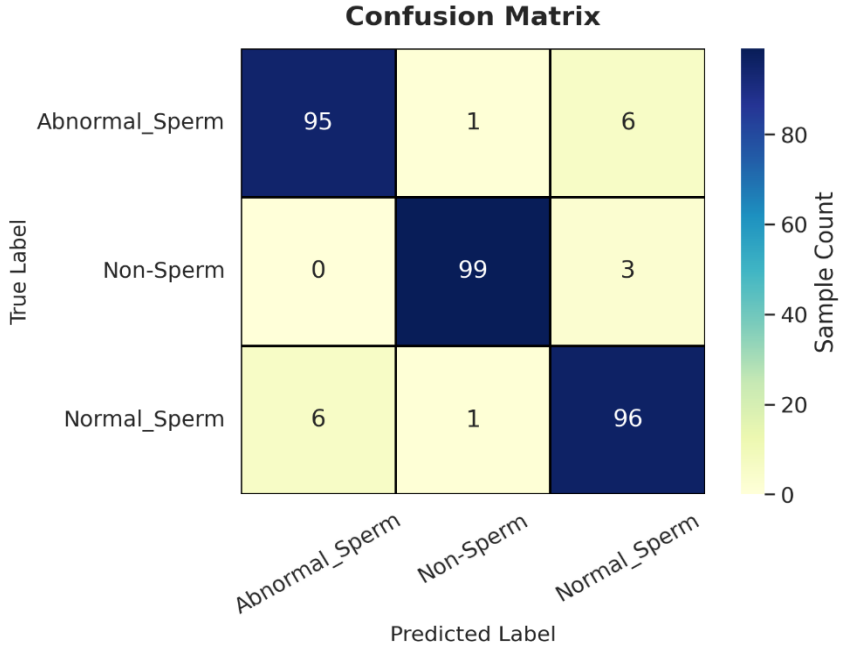

FIGURE: 8 CONFUSION MATRIX

Importantly, the non-sperm category exhibited near-perfect classification, with only 3 misclassifications out of 102, indicating the model's strong capacity to distinguish between sperm-containing and non-sperm images. This is crucial for ensuring that non-relevant samples are effectively filtered out in clinical and laboratory workflows. The low false positive and false

negative rates across all classes reinforce the model's high precision and recall, which are essential for real-world applicability. From a diagnostic standpoint, the ability to accurately differentiate abnormal from normal sperm morphology has direct implications for fertility assessments and andrology research. The minor confusion between Normal and Abnormal sperm likely arises from subtle morphological similarities or image quality variability, which are common challenges in biomedical image interpretation.

Overall, the confusion matrix in Figure X demonstrates the robustness and clinical readiness of the proposed architecture. The model not only generalizes well across diverse sperm morphology types but also maintains a balanced sensitivity and specificity profile. This balance is critical in ensuring diagnostic reliability, minimizing misclassification-induced intervention errors, and ultimately supporting informed decision-making in reproductive health settings.

ROC–AUC AND PRECISION–RECALL ANALYSIS

To thoroughly assess the discriminative power of the proposed CBAM-enhanced Xception model across all sperm morphology classes, both Receiver Operating Characteristic and Precision–Recall curves were also performed to fully understand the discriminative power of the proposed CBAM-enhanced Xception model amongst all sperm morphology classes. In Figure 9, the ROC curves for all three classes of Abnormal Sperm (class 0), non-sperm (class 1), and Normal Sperm (class 2) are shown to have high area under curve (AUC) values of 0.99 for all the classes which show high sensitivity and specificity. The ROC curves all near the ideal top-left corner of any ROC space meaning that the model is successful in minimizing false negative and false positives.

The PR curves shown in Figure 10 help add further validation to the model's robustness especially under class imbalance conditions. The Average Precision (AP) for class 0, class 1, and class 2 were 0.98, 0.99, and 0.96 respectively, which validates the model as producing true positives while avoiding false alarms and is indicative of a high level of precision in the model. The shapes of the PR curves, particularly their steep incline, the plateau avoided the upper-right region where both recall and precision are present than that deep dive into each class proves the model retains precision even at high recall values.

This factor is necessary in regard to deploying the model in medical diagnostic settings as misclassifying samples could result in dire consequences. Taken together, these curve-based evaluations provide a comprehensive validation of the model's discriminative capacity, showing strong alignment between predicted and true classes across both global (ROC) and class-

sensitive (PR) performance metrics. This ensures balanced classification, excellent class-wise reliability, and strong clinical relevance for the deployment of the model in real-world reproductive diagnostics.

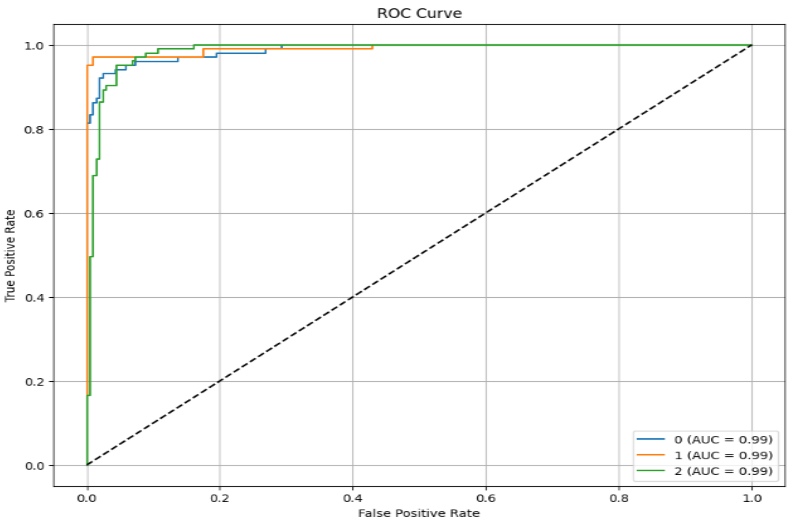

FIGURE:9 ROC CURVE OF THE PROPOSED MODEL

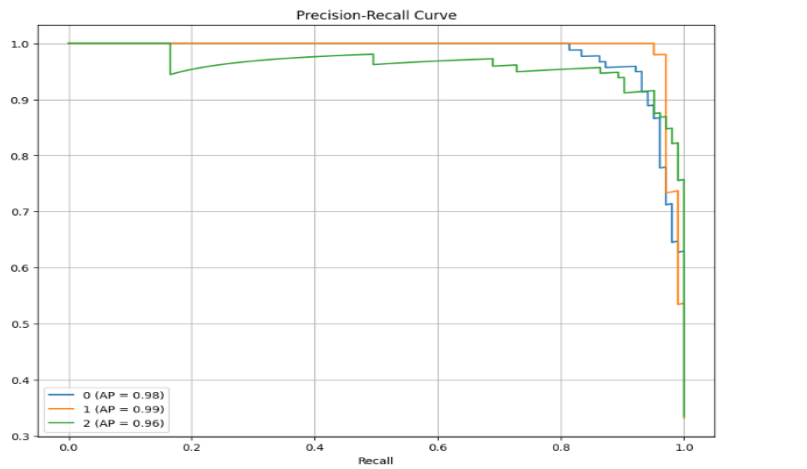

FIGURE:10 PRECISION-RECALL CURVE

CLASS-WISE PERFORMANCE EVALUATION

To comprehensively evaluate the performance of the proposed model across different sperm morphology categories, we analyzed precision, recall, and F1-score metrics per class, as illustrated in Figure 11. These metrics provide a nuanced view of the model's classification effectiveness, highlighting its strengths and possible class-specific limitations.

The model scored a precision of 0.94, recall of 0.98, and F1-score of 0.91 when identifying the Abnormal Sperm class, demonstrating good sensitivity when locating abnormal samples with some false positives in this class. For non-sperm class, the model produced balanced performance metrics of a precision and F1-score of 0.93 with a recall of 0.97 indicating trustworthy recognitions with minimal misclassifications. For the Normal Sperm class, the model recorded a precision and a recall of 0.94 and 0.98 respectively, with an F1- score of 0.95 also suggesting consistent and confident predictions in this class as well.

The narrow range of scores demonstrates consistency that allows the model to provide similar predictive quality across all classes, therefore avoiding the over-generalization or overfitting towards a particular class. This consistency is important for biomedical classification tasks as imbalanced or biased recognition can materially impact clinical decisions. The high F1-scores also validate that the model maintains a balanced trade-off between precision and recall, albeit, for sperm morphology classification problems, the complexities, and variability of the final SPM predictably implies the model can confidently accommodate this.

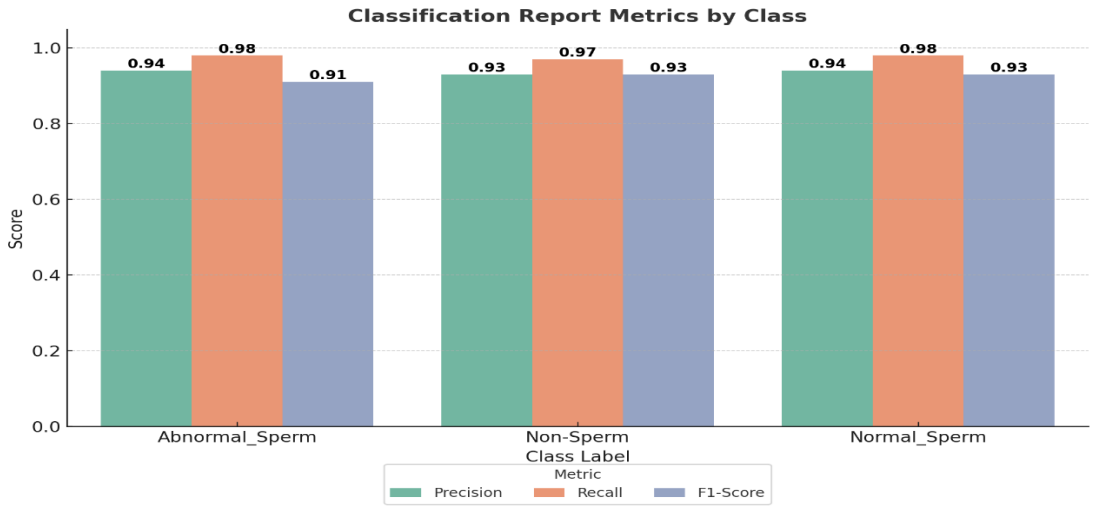

FIGURE:11 CLASSIFICATION REPORT OF PROPOSED METHOD

QUALITATIVE ASSESSMENT OF MODEL PREDICTIONS

Figure 12 showcases a selection of test samples to qualitatively evaluate the classification performance of the proposed model across the three defined categories: Class 0 – Abnormal Sperm, Class 1 Non-Sperm, and Class 2 Normal Sperm. Each sub-image includes the ground truth label and the corresponding model prediction, with correct classifications annotated in green and misclassifications in red.

The visual results demonstrate the model's strong discriminative capability in identifying subtle morphological features that distinguish between abnormal, normal, and non-sperm cells. In particular, the model consistently recognized class 1 (non-sperm), benefiting from its distinct structural characteristics. Similarly, classes 0 and 2 were also correctly classified in most cases, highlighting the effectiveness of the model's learned spatial and contextual representations.

The few observed misclassifications, predominantly involving confusion between classes 0 and 2, are attributable to overlapping morphological traits, ambiguous boundaries, or variations in staining quality and contrast. These challenges reflect common limitations in microscopic sperm morphology imaging and underscore the importance of robust feature learning. Overall, this qualitative inspection complements the quantitative results by illustrating the model's predictive behavior on real samples and confirming its capacity to generalize well across diverse image conditions.

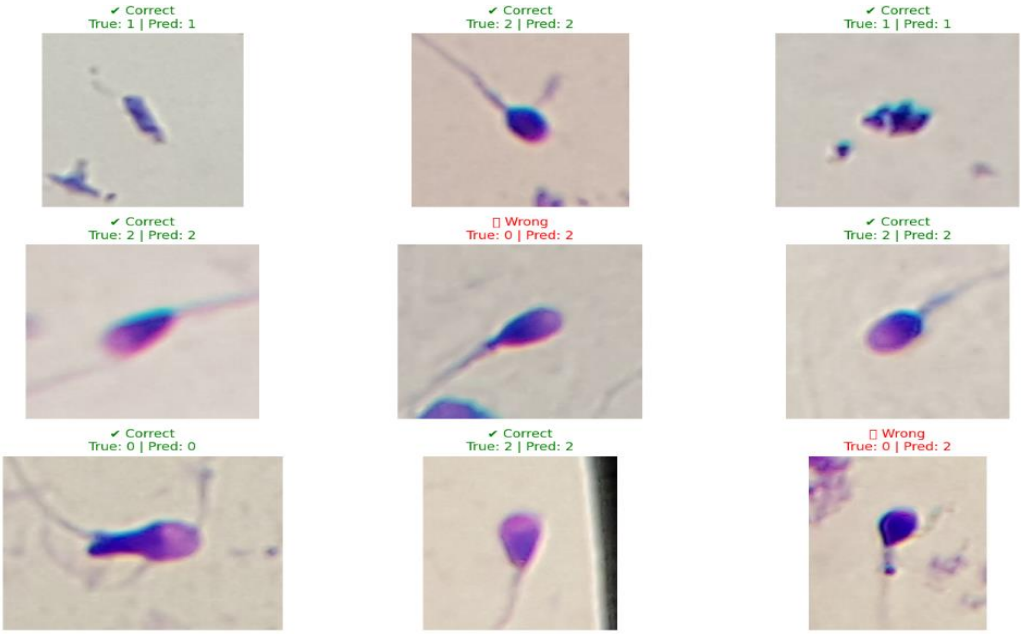

FIGURE:12 SAMPLE PREDICTIONS ACROSS SPERM MORPHOLOGY CLASSES

CONCLUSION AND FUTURE WORK

This study provided a framework that used Xception architecture with the Convolutional Block Attention Module, and it reported the framework performance on the SMIDS dataset. The proposed model achieved good results test accuracy of 96.2%, a macro-averaged F1 score of 95.1 and an AUC of 0.99 which meant there was high maximum-accuracy ability of the model to discriminate the Normal, Abnormal, non-Sperm categories, With the help of CBAM's channel and

spatial attention, the model demonstrated its ability to extract attention from the diagnostically relevant morphologically structures and suppress background noise. By utilizing CBAM's attention mechanism, the model was able to address the inherent challenges of traditional CNNs and the fine-grained, subtle differences in medical sperm morphology images. The robustness and clinical relevance of the framework's model were evaluated by multiple metrics (i.e., confusion matrices, ROC curves, precision-recall analysis, and reporting performance metrics across classes). It was shown that evaluating across these measures demonstrated the generalizability of the proposed framework overall. The results of the framework bode well because they suggest that the Xception-CBAM architecture presented is scalable, interpretable, and efficient and exemplifies a tool for supporting automated fertility diagnostics in clinical and laboratory settings.

Despite the promising outcomes, several areas for further investigation remain. The dataset used, though balanced and well-curated, is relatively small when compared to the data volumes encountered in real clinical environments. Future studies should focus on expanding the training data through the collection of larger, multi-center datasets that represent greater biological and technical diversity. Furthermore, the current study relied solely on static image inputs. Incorporating clinical metadata such as patient history, hormone levels, or motility parameters may enhance model performance and provide a more holistic fertility analysis. In addition, the framework could be extended by exploring advanced deep learning architectures, including transformer-based or hybrid attention models, which may capture long-range dependencies and further boost classification accuracy. Real-time deployment, particularly through lightweight, mobile-compatible versions of the model, could facilitate point-of-care diagnostics in resource-limited settings. Finally, increasing the interpretability of model predictions through explainable AI techniques, such as Grad-CAM or SHAP, would enhance clinical trust and adoption by allowing practitioners to visualize and understand the decision-making process behind each classification.

REFERENCES

[1] D. L. Pelzman and J. I. Sandlow, "Sperm morphology: Evaluating its clinical relevance in contemporary fertility practice," *Reproductive Medicine and Biology*, vol. 23, no. 1, Jan. 2024, doi: https://doi.org/10.1002/rmb2.12594.

[2] R. Menkveld, Cas A G Holleboom, and Johann P T Rhemrev, "Measurement and significance of sperm morphology," *Asian Journal of Andrology*, vol. 13, no. 1, 2011, doi: https://doi.org/10.1038/aja.2010.67.

[3] F. Boitrelle *et al.*, "The Sixth Edition of the WHO Manual for Human Semen Analysis: A Critical Review and SWOT Analysis," *Life*, vol. 11, no. 12, p. 1368, Dec. 2021, Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8706130/

[4] Kowsar Qaderi, Foruzan Sharifipour, Mahsa Dabir, R. Shams, and A. Behmanesh, "Artificial intelligence (AI) approaches to male infertility in IVF: a mapping review," *European journal of medical research*, vol. 30, no. 1, Apr. 2025, doi: https://doi.org/10.1186/s40001-025-02479-6.

[5] Tariq Rahim, Ibadullah, Aimal Nazir, Muhammad Salih Tanveer, and Muhammad Rohan Qureshi, "A DEEP LEARNING APPROACH TO PCOS DIAGNOSIS: TWO-STREAM CNN WITH TRANSFORMER ATTENTION MECHANISM", *SES*, vol. 3, no. 7, pp. 1–20, Jul. 2025

[6] Ibomoiye Domor Mienye, T. G. Swart, G. Obaido, M. Jordan, and P. Ilono, "Deep Convolutional Neural Networks in Medical Image Analysis: A Review," *Information*, vol. 16, no. 3, p. 195, Mar. 2025, doi: https://doi.org/10.3390/info16030195.

[7] A. Agarwal *et al.*, "Sperm Morphology Assessment in the Era of Intracytoplasmic Sperm Injection: Reliable Results Require Focus on Standardization, Quality Control, and Training," *The World Journal of Men's Health*, vol. 40, no. 3, p. 347, 2022, doi: https://doi.org/10.5534/wjmh.210054.

[8] A. Gongora, "A new perspective on sperm analysis through artificial intelligence: The path toward personalized reproductive medicine," *Ann. Clin. Med. Case Rep.*, vol. 14, no. 11, pp. 1–10, 2025.

[9] I. Iqbal, G. Mustafa, and J. Ma, "Deep Learning-Based Morphological Classification of Human Sperm Heads," *Diagnostics*, vol. 10, no. 5, p. 325, May 2020, doi: https://doi.org/10.3390/diagnostics10050325.

[10] A. G. Pradnya Sidhawara, "Male Fertility Classification using Machine Learning and Oversampling Techniques," *Jurnal Buana Informatika*, vol. 15, no. 01, pp. 1–10, Apr. 2024, doi: https://doi.org/10.24002/jbi.v15i1.8718.

[11] H. O. Ilhan, G. Serbes, and N. Aydin, "Automated sperm morphology analysis approach using a directional masking technique," *Computers in Biology and Medicine*, vol. 122, p. 103845, Jul. 2020, doi: https://doi.org/10.1016/j.compbiomed.2020.103845.

[12] Aristoteles Aristoteles, Ridho Sholehurrohman, and Nasywa Nathania Wirawan, "Human

Sperm Morphology Classification Using YOLOv5 Deep Learning Algorithm," *International Journal of Electronics and Communications Systems*, vol. 4, no. 2, pp. 99–99, Dec. 2024, doi: https://doi.org/10.24042/ijecs.v4i2.24419.

[13] J. Riordon, C. McCallum, and D. Sinton, "Deep learning for the classification of human sperm," *Computers in Biology and Medicine*, vol. 111, pp. 103342–103342, Aug. 2019, doi: https://doi.org/10.1016/j.compbiomed.2019.103342.

[14] S. Shahzad, M. Ilyas, M. Ikram, Hafiz Tayyab Rauf, Seifedine Kadry, and Emad Abouel Nasr, "Sperm Abnormality Detection Using Sequential Deep Neural Network," *Mathematics*, vol. 11, no. 3, pp. 515–515, Jan. 2023, doi: https://doi.org/10.3390/math11030515.

[15] H. Jabbari and Nooshin Bigdeli, "New conditional generative adversarial capsule network for imbalanced classification of human sperm head images," *Neural Computing and Applications*, vol. 35, no. 27, pp. 19919–19934, Jul. 2023, doi: https://doi.org/10.1007/s00521-023-08742-3.

[16] H. O. Ilhan, I. O. Sigirci, G. Serbes, and N. Aydin, "A fully automated hybrid human sperm detection and classification system based on mobile-net and the performance comparison with conventional methods," *Medical & Biological Engineering & Computing*, vol. 58, no. 5, pp. 1047–1068, Mar. 2020, doi: https://doi.org/10.1007/s11517-019-02101-y.

[17] Kowsar Qaderi, Foruzan Sharifipour, Mahsa Dabir, R. Shams, and A. Behmanesh, "Artificial intelligence (AI) approaches to male infertility in IVF: a mapping review," *European journal of medical research*, vol. 30, no. 1, Apr. 2025, doi: https://doi.org/10.1186/s40001-025-02479-6.

[18] S. A. Hicks *et al.*, "Machine Learning-Based Analysis of Sperm Videos and Participant Data for Male Fertility Prediction," *Scientific Reports*, vol. 9, no. 1, Nov. 2019, doi: https://doi.org/10.1038/s41598-019-53217-y.

[19] K. R. Seymour, J. Saam, K. R. Pool, T. Pini, J. P. Rickard, and P. de, "Accurate Classification of Ram Sperm Morphology into 2 and 5 Categories Using an Objective Machine Learning Model," Jan. 2025, doi: https://doi.org/10.2139/ssrn.5333984.

[20] J. B. You, C. McCallum, Y. Wang, J. Riordon, R. Nosrati, and D. Sinton, "Machine learning for sperm selection," *Nature Reviews Urology*, May 2021, doi: https://doi.org/10.1038/s41585-021-00465-1.

[21] Orvile, "Sperm Morphology Image Data Set (SMIDS)," *Kaggle.com*, 2022. https://www.kaggle.com/datasets/orvile/sperm-morphology-image-data-set-smids.

[22] M. A. Rahman, M. Badrul, M. A. Hossain, and A. S. M. Sanwar Hosen, "Enhanced Brain Tumor Classification Using MobileNetV2: A Comprehensive Preprocessing and Fine-Tuning Approach," *BioMedInformatics*, vol. 5, no. 2, pp. 30–30, Jun. 2025, doi: https://doi.org/10.3390/biomedinformatics5020030.

[23] J. Brownlee, "Random Oversampling and Undersampling for Imbalanced Classification," *Machine Learning Mastery*, Jan. 14, 2020. https://machinelearningmastery.com/random-oversampling-and-undersampling-for-imbalanced-classification/

[24] Seyed Iman Saedi and M. Rezaei, "A Modified Xception Deep Learning Model for Automatic Sorting of Olives Based on Ripening Stages," *Inventions*, vol. 9, no. 1, pp. 6–6, Dec. 2023, doi: https://doi.org/10.3390/inventions9010006.

[25] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional Block Attention Module," *openaccess.thecvf.com*, 2018. https://openaccess.thecvf.com/content_ECCV_2018/html/Sanghyun_Woo_Convolutional_Block_Attention_ECCV_2018_paper.html

[26] S. Ibrahim *et al.*, "Improving performance evaluation coefficient and parabolic solar collector efficiency with hybrid nanofluid by innovative slotted turbulators," *Sustainable Energy Technologies and Assessments*, vol. 53, p. 102391, Oct. 2022, doi: https://doi.org/10.1016/j.seta.2022.102391.

[27] M. Z. Alom *et al.*, "A State-of-the-Art Survey on Deep Learning Theory and Architectures," *Electronics*, vol. 8, no. 3, p. 292, Mar. 2019, doi: https://doi.org/10.3390/electronics8030292.

[28] S. Ibrahim, D. N. Khan Marwat, N. Ullah, and K. S. Nisar, "Investigation of fluid flow pattern in a 3D meandering tube," *Frontiers in Materials*, vol. 10, Jun. 2023, doi: https://doi.org/10.3389/fmats.2023.1187986.

[29] Kaggle, "Kaggle: Your Machine Learning and Data Science Community," *Kaggle*, [Online]. Available: https://www.kaggle.com/code.