West Visayas State University
(Formerly Iloilo Normal School)
COLLEGE OF INFORMATION AND COMMUNICATIONS TECHNOLOGY
Luna St., La Paz, Iloilo City 5000
Iloilo, Philippines
* Trunkline: (063) (033) 320-0870 to 78 loc. 1403 * Telefax No.: (033) 320-0879
* Website: www.wvsu.edu.ph * Email Address: cict@wvsu.edu.ph

Group Members:	Course/Year/Section:	Date:
Joeross Palabica Melissa Marielle Valdez Leann Villaruel	BSCS 3B - AI	09/12/2025

Sperm Morphology Classification Documentation

I. Problem Statement

Checking whether sperm cells look normal is a major challenge in fertility diagnostics. Today, embryologists manually evaluate sperm cells under a microscope to determine if their morphology is normal or abnormal. However, manual assessment is highly subjective, inconsistent, and prone to inter- and intra-observer variability (Finelli et al., 2021). Variations in interpretation lead to inconsistent fertility diagnoses and treatment decisions. Sperm morphology assessment focuses on evaluating structural features such as head shape, acrosome structure, midpiece characteristics, and tail abnormalities

Because these features are subtle and require precise evaluation, human judgment alone often results in unreliable outcomes.

II. Current Solutions

Computer-Assisted Sperm Analysis (CASA) systems have been used in fertility clinics for over two decades (Belala et al., 2024). They automate the evaluation of sperm concentration, motility, and morphology using classical image-processing techniques (Valverde et al., 2020).

While CASA performs well in motility and concentration measurement, it performs poorly in morphology classification due to highly complex sperm structures, necessity of manual feature extraction, and limited generalization across clinics. Studies show that CASA systems inconsistently classify sperm morphology, making them insufficient for robust clinical decision-making (Finelli et al., 2021).

III. Dataset Information

This project uses the Sperm Morphology Image Data Set (SMIDS) dataset, publicly available on Kaggle. SMIDS was collected using a smartphone-based data acquisition method that was originally developed for detecting and counting motile sperm cells. This method was later validated for sperm concentration analysis, showing a strong correlation with manual counting results.

Unlike other sperm morphology datasets, SMIDS does not divide abnormal samples into specific subcategories. Instead, sperm cells are labeled simply as normal (1021), abnormal (1005), and non-sperm (974). All images are stored in RGB format, and each image may include noise, multiple sperm heads, or overlapping tails.

The dataset is publicly accessible through the following link:

https://www.kaggle.com/datasets/orvile/sperm-morphology-image-data-set-smids

IV. Solution Overview

A VGG11-based Convolutional Neural Network (CNN) was **developed and trained from scratch** to classify sperm morphology into three categories. The model was modified to accept 1-channel grayscale inputs and does not use any pretrained weights.

In this study, the network learns morphological patterns directly from microscopy images without relying on handcrafted features. This allows the model to capture subtle variations in sperm head and tail structure, even in the presence of noise or overlapping cells.

Several techniques were integrated to improve model performance and reduce overfitting:

- ***Data augmentation*** including rotation, horizontal flipping, affine transformations, and brightness/contrast jitter
- ***Input normalization*** using dataset-specific statistics (mean = 0.7303, std = 0.1039)
- ***Weighted Cross-Entropy Loss*** with label smoothing (0.1) to address class imbalance
- ***Early stopping*** to halt training when validation accuracy no longer improves
- ***Cosine annealing learning-rate scheduling*** for smoother optimization
- ***Hyperparameter*** sweep to tune the optimizer, learning rate, and weight decay

V. Network Structure

The model uses a VGG11-style Convolutional Neural Network adapted for 1-channel grayscale input at 112×112. Its feature extractor consists of sequential convolutional layers with 3×3 kernels, each followed by ReLU activations, and periodic 2×2 max-pooling to downsample. There are no batch normalization layers inside the network. The classifier head has three fully connected layers: 4096 → 4096 → 3, with dropout in the classifier to improve generalization. The final layer outputs logits for three classes (Normal_Sperm, Abnormal_Sperm, Non-Sperm). Total trainable parameters are approximately 128.8 million, reflecting a high-capacity architecture suitable for learning complex morphological patterns.

Input Layer

- The model processes 112×112 grayscale microscopy images as input. Each image is converted into a tensor with the shape 1 × 112 × 112 (channels × height × width) to align with the modified VGG11 architecture.
- Before entering the network, images undergo several preprocessing steps: they are converted to grayscale, resized, center-cropped to 112×112, and standardized using dataset-specific normalization values (mean = 0.7303, standard deviation = 0.1039).
- During both training and inference, normalization is consistently applied through transforms.Normalize ([0.7303], [0.1039]) to ensure the input distribution remains aligned with the conditions under which the model was trained.
-

Convolutional Blocks

The network is composed of a series of convolutional blocks designed to extract increasingly complex morphological features from the sperm images. Each block applies a 3×3 convolution to capture spatial patterns such as edges, contours, and fine structural details, followed by a ReLU activation to introduce non-linearity. Periodic 2×2 max-pooling layers downsample the feature maps, expanding the receptive field while reducing computational complexity.

No batch normalization layers are included in this architecture; normalization is handled exclusively at the input stage through preprocessing.

As images propagate through deeper layers, the receptive field gradually expands. Early layers focus on low-level visual cues such as edges and intensity gradients. Mid-level layers begin to identify morphological components like head curvature, vacuole positions, and tail orientation. The deepest layers integrate these features into high-level representations that help distinguish between normal and abnormal sperm morphology.

Feature Extraction Stage

The image passes through VGG11 convolutional and max-pooling layers, producing deep feature maps. For 112×112 input, the final feature map is approximately 512 × 3 × 3.

These maps are flattened into a 4,608-dimensional vector, which compactly encodes salient morphological cues (e.g., head shape, vacuole position, tail orientation) learned by the network. This vector is then fed to the classifier (4096 → 4096 → 3) to produce the final logits.

Fully Connected Layers

The flattened feature vector (≈4,608 dims from 512×3×3) is fed into VGG11's classifier: 4096 → 4096 → 3. Each dense layer uses ReLU activations; dropout is applied in the classifier to reduce overfitting.

This stage maps learned features to class logits, progressively compressing the representation toward the three target categories (Normal_Sperm, Abnormal_Sperm, Non-Sperm).

Output Layer and Loss

The network's final layer is a linear classifier that produces three logits (one per class: Normal_Sperm, Abnormal_Sperm, Non-Sperm).

During inference, a softmax is applied to these logits to obtain class probabilities, and the highest-probability class is reported as the prediction.

Training uses Weighted CrossEntropyLoss (with label smoothing=0.1), which consumes logits directly and applies class weights derived from the training set's class frequencies to mitigate imbalance.

Training Environment

- Training runs on a CUDA GPU.
  Gradients are computed via backpropagation to update convolutional filters and classifier weights across ~128.8M parameters.
- Optimization with cosine annealing learning-rate scheduling.
- Loss: Weighted CrossEntropyLoss with label smoothing=0.1 to mitigate class imbalance.

- ➢ Regularization: data augmentation (rotation, flip, affine, brightness/contrast jitter) and dropout in the VGG11 classifier.
- ➢ Early stopping saves the best checkpoint (vgg11_model_L.pt) when validation accuracy improves, preventing unnecessary epochs.

This setup enables the CNN to learn morphology directly from normalized grayscale images without handcrafted features.

VI. Tools and Development Environment

The model development was performed using Python and PyTorch. Data loading, preprocessing, and augmentation were implemented using TorchVision utilities. GPU support provided by CUDA was essential for training the deep CNN efficiently. Additional Python libraries were used for numerical computations and visualization of training curves when needed.

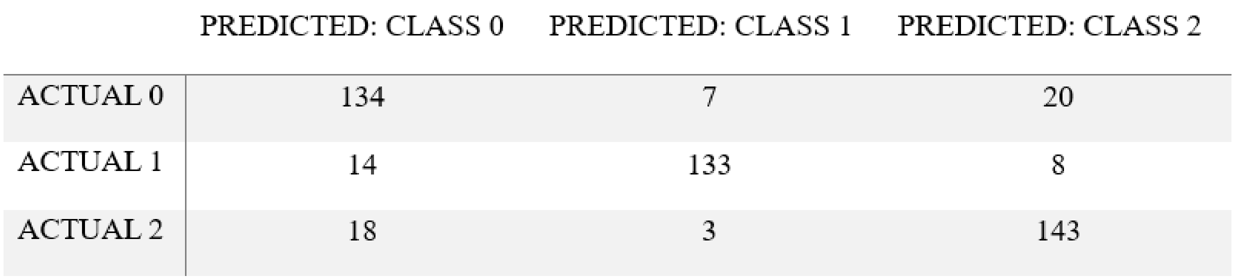
The dataset was sourced from Kaggle and processed locally to compute normalization parameters such as mean and standard deviation. These values were applied during preprocessing to stabilize training. Weighted cross-entropy loss was employed to mitigate class imbalance, with class weights computed directly from the dataset distribution.

VII. Validation Performance

The model's performance was first evaluated on the validation set to monitor generalization during development. The results indicate strong discriminatory capability across the three morphology classes.

- ➢ **Validation Loss:** 0.391
- ➢ **Validation Accuracy:** 85.42%
- ➢ **Validation F1 Score:** 0.855
- ➢ **Validation ROC AUC:** 0.958

Validation Confusion Matrix



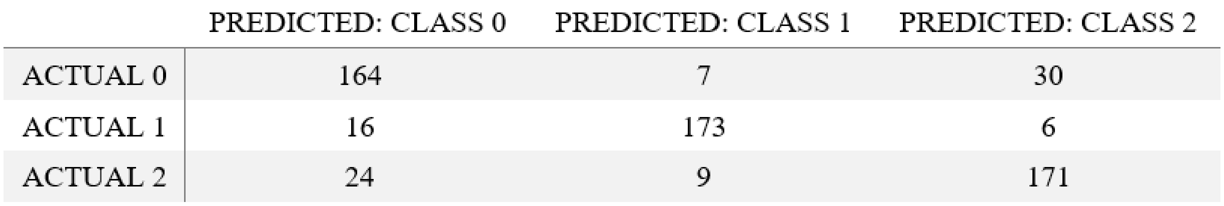
	PREDICTED: CLASS 0	PREDICTED: CLASS 1	PREDICTED: CLASS 2
ACTUAL 0	134	7	20
ACTUAL 1	14	133	8
ACTUAL 2	18	3	143

VIII. Test Performance

The final evaluation on the held-out test set shows that the model generalizes well, with metrics closely mirroring the validation performance.

- **Test Loss:** 0.398
- **Test Accuracy:** 84.67%
- **Test F1 Score:** 0.848
- **Test ROC AUC:** 0.959

Test Confusion Matrix



	PREDICTED: CLASS 0	PREDICTED: CLASS 1	PREDICTED: CLASS 2
ACTUAL 0	164	7	30
ACTUAL 1	16	173	6
ACTUAL 2	24	9	171

IX. Hyperparameter Tuning

Hyperparameter tuning was conducted to optimize the model’s learning dynamics and improve overall classification performance. A short exploratory sweep was performed to identify suitable training configurations before full training.

A 3-epoch preliminary sweep was executed on the validation split using the following search space:

- Optimizer: Adam, AdamW, or SGD
- Learning rate: 1×10^{-4}, 3×10^{-4}, or 1×10^{-3}
- Weight decay: 0.0, 1×10^{-4}, or 5×10^{-4}

Model variants were compared based on macro-F1 score, with accuracy used as a secondary criterion in the event of ties.

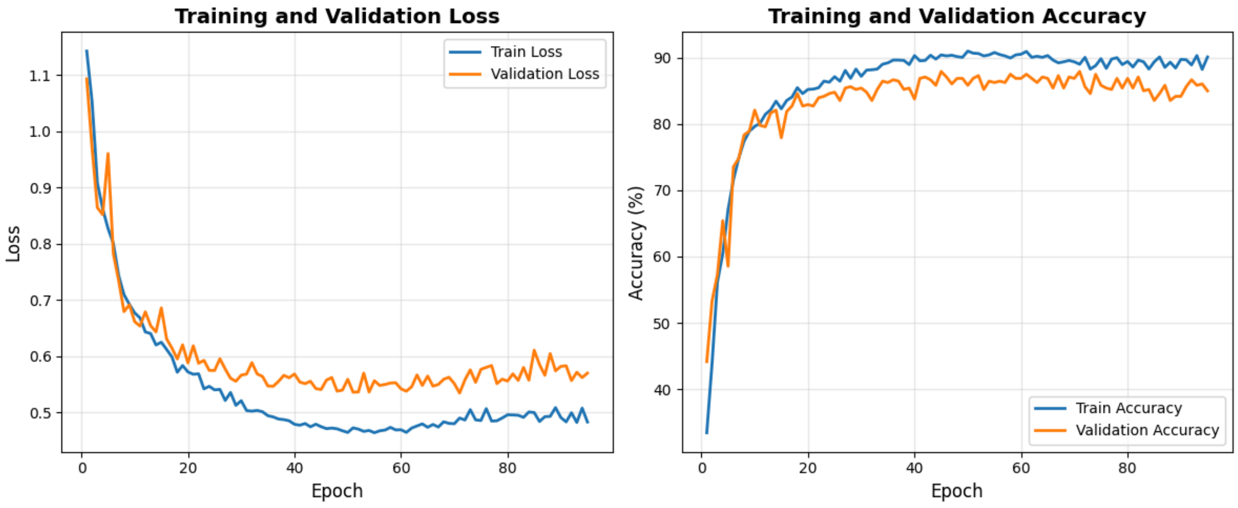
The best-performing configuration selected for full training used:

- Optimizer: AdamW
- Initial learning rate: 0.0001
- Weight decay: 0.0
- Loss function: Cross-Entropy with class weights and label smoothing (0.1)
- Batch size: 32
- Learning-rate scheduler: CosineAnnealingLR (T_max = 50)
- Training schedule: Up to 100 epochs with early stopping (patience = 50)

The final model weights were stored in the checkpoint file vgg11_model_L.pt.

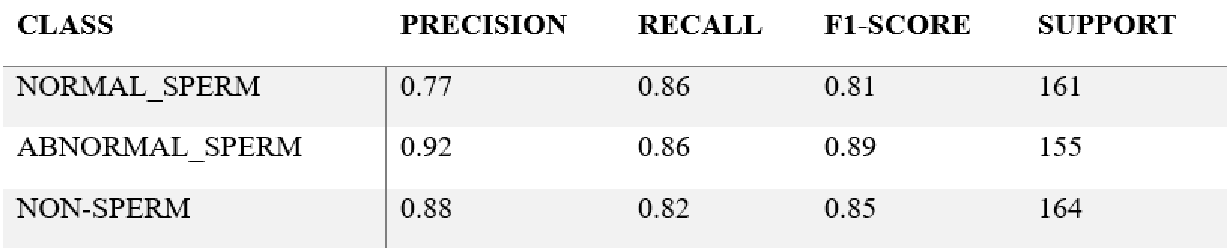
X. Training and Validation Performance Curves

The following plots illustrate the progression of the model’s performance throughout training. The training and validation loss curves show how the model converged over successive epochs and provide insight into possible overfitting or underfitting. Likewise, the training and validation accuracy curves track performance improvements and reflect how well the model generalized to unseen data during training.



XI. Classification Report on Test Set

The model's performance was evaluated on the held-out test set using standard classification metrics: precision, recall, and F1-score. The results for each class and overall averages are summarized below:



CLASS	PRECISION	RECALL	F1-SCORE	SUPPORT
NORMAL_SPERM	0.77	0.86	0.81	161
ABNORMAL_SPERM	0.92	0.86	0.89	155
NON-SPERM	0.88	0.82	0.85	164

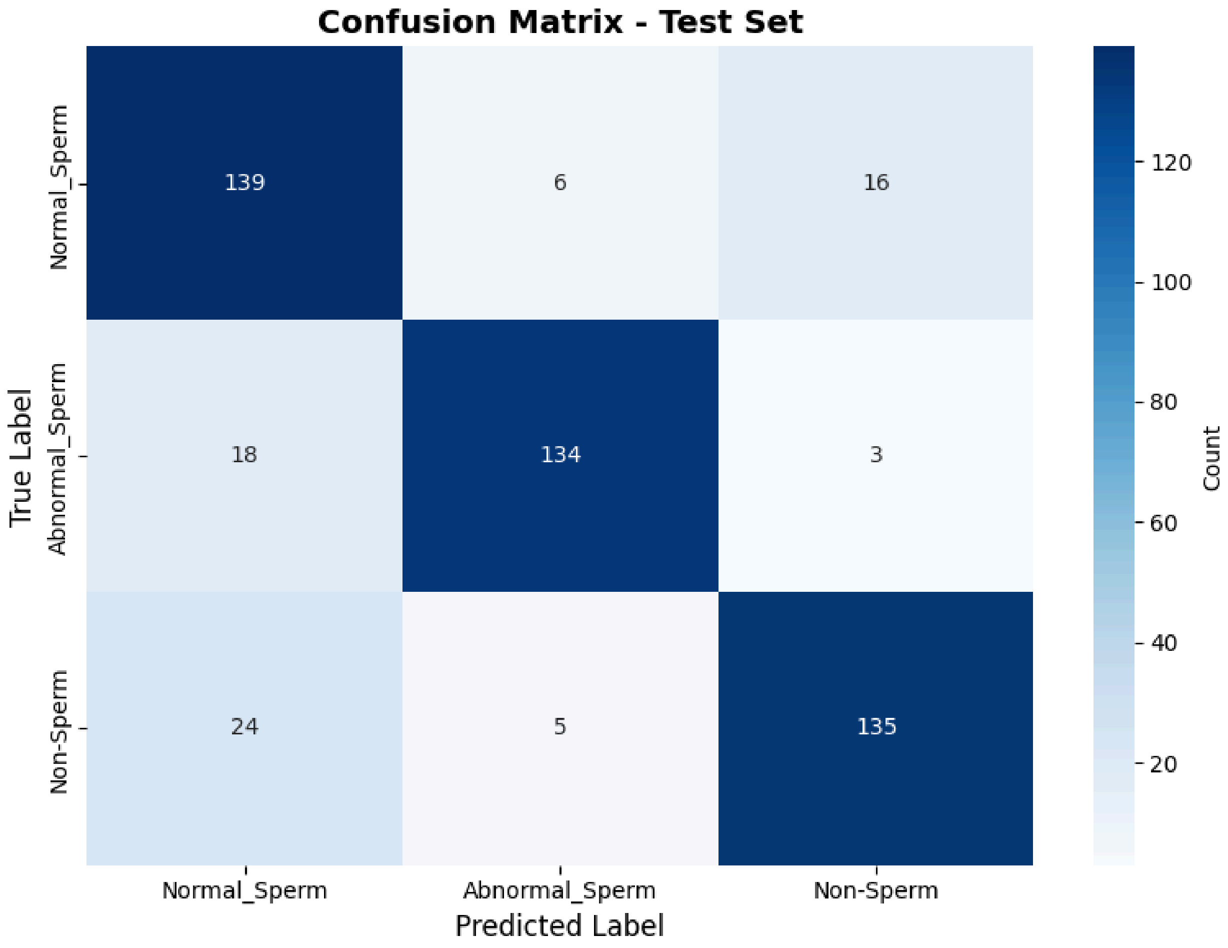
Overall performance:

- ➢ Accuracy: 0.85
- ➢ Macro average: Precision = 0.86, Recall = 0.85, F1-score = 0.85
- ➢ Weighted average: Precision = 0.86, Recall = 0.85, F1-score = 0.85

The results indicate that the model performs well across all classes, with particularly high precision for Abnormal_Sperm and balanced performance across normal, abnormal, and non-sperm categories.

XII. Confusion Matrix

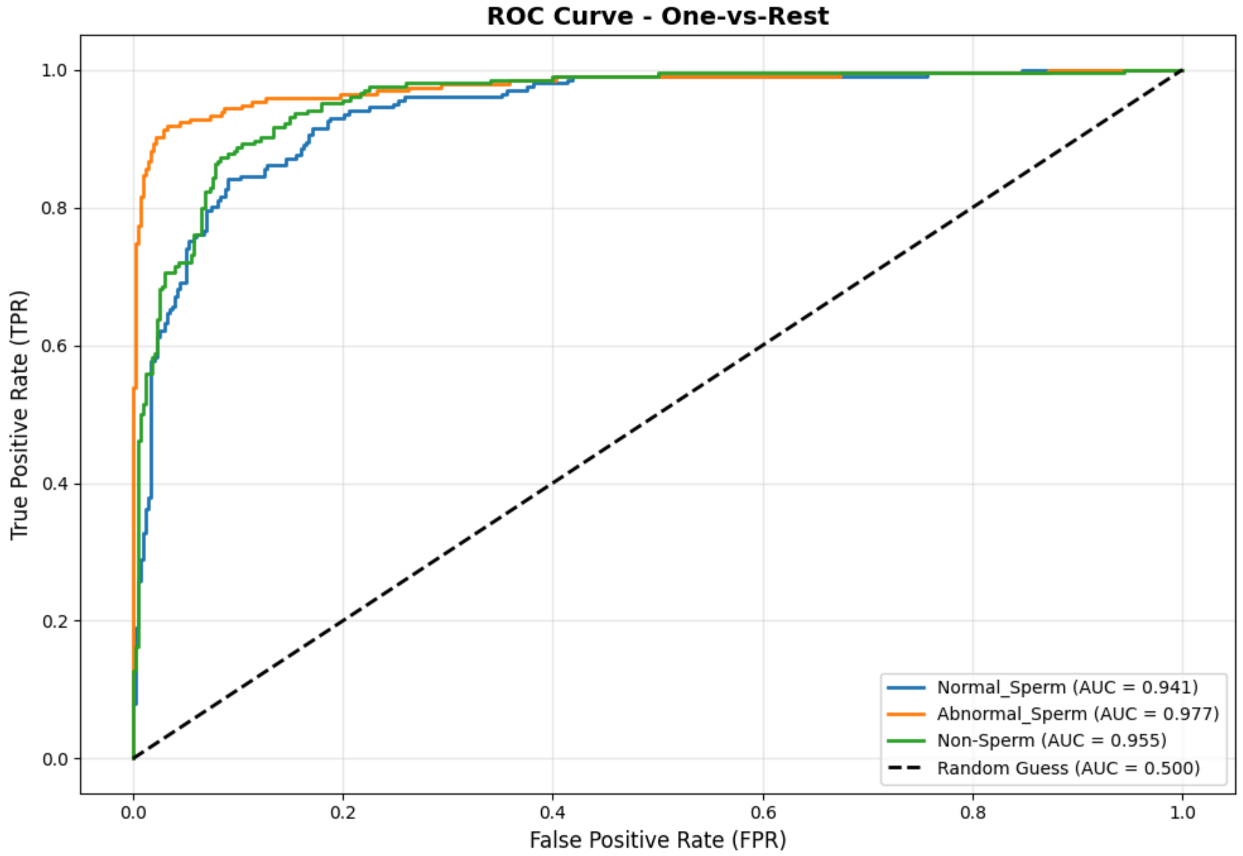
The confusion matrix summarizes predicted versus true labels for the three classes, making class-specific errors visible. It highlights where the classifier confuses Normal_Sperm, Abnormal_Sperm, and Non-Sperm, providing a clearer view of per-class performance than overall accuracy alone.



Interpretation: a strong diagonal indicates correct predictions, while larger off-diagonal cells reveal systematic confusions. If misclassifications cluster between specific classes, consider targeted augmentation, class-weight tuning, threshold adjustments, or preprocessing refinements.

XIII. ROC Curve Analysis

The ROC curves (one-vs-rest) visualize the trade-off between true positive rate and false positive rate for each class, showing how well the model separates Normal_Sperm, Abnormal_Sperm, and Non-Sperm across all thresholds. This complements accuracy by highlighting threshold-independent discrimination.

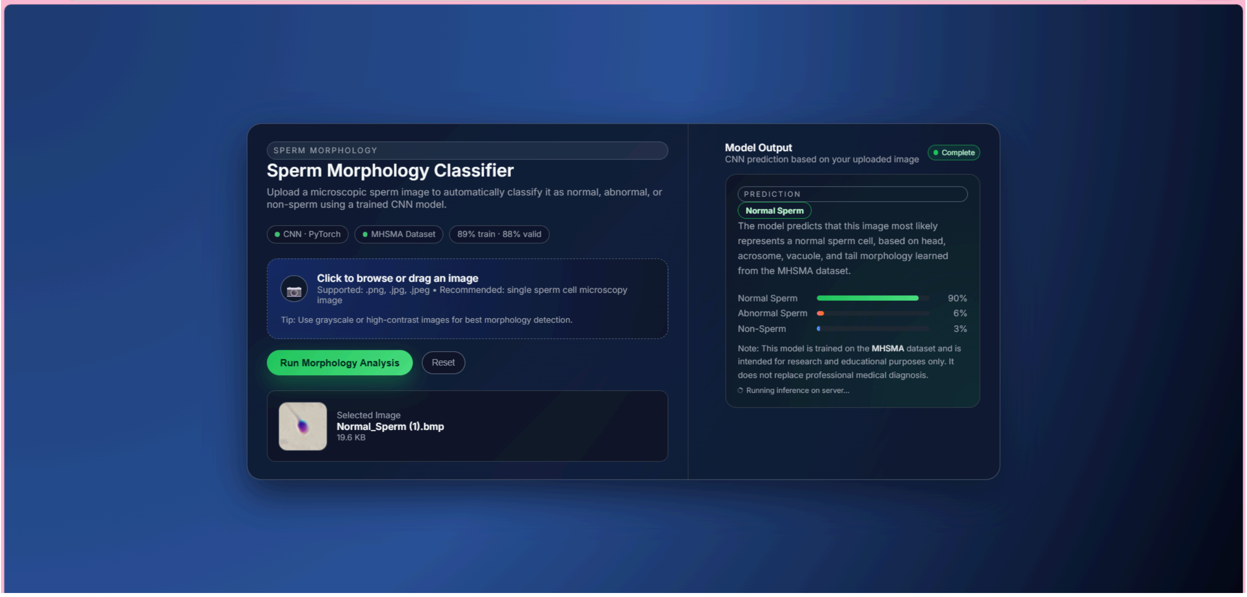


Interpretation: curves nearer the top-left and higher AUC values indicate stronger separability. Classes with lower AUC or curves closer to the diagonal may benefit from more data, targeted augmentation, class-weight tuning, or threshold adjustments.

XIV. Web Application

The model is deployed via a Flask web API with an HTML/JavaScript front end. The model is loaded once for efficient inference, and uploaded images are automatically preprocessed (grayscale conversion, resizing to 112×112, and normalization using training statistics). The

/predict endpoint outputs the predicted class and probabilities for Normal, Abnormal, and Non-Sperm. Users can upload images via drag-and-drop, preview them in real time, and run the analysis with a single button, with results displayed clearly on the interface.



Sperm Morphology Classifier

Upload a microscopic sperm image to automatically classify it as normal, abnormal, or non-sperm using a trained CNN model.

Click to browse or drag an image
Supported: .png, .jpg, .jpeg • Recommended: single sperm cell microscopy image

Tip: Use grayscale or high-contrast images for best morphology detection.

Selected Image
Normal_Sperm (1).bmp
19.6 KB

Model Output
CNN prediction based on your uploaded image

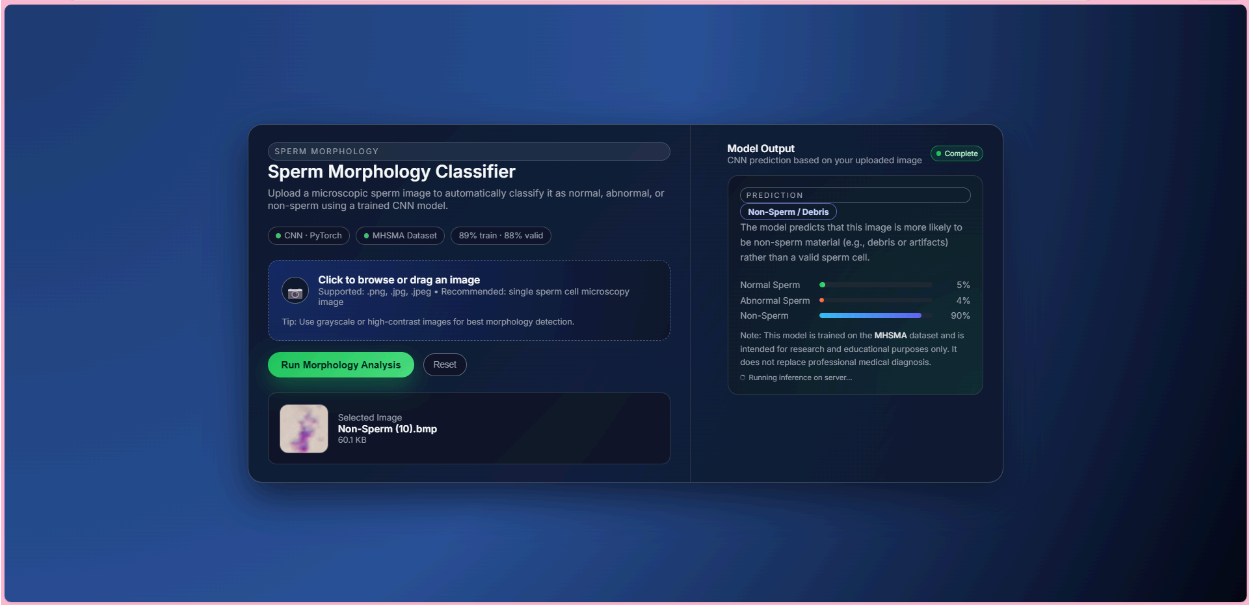
The model predicts that this image most likely represents a normal sperm cell, based on head, acrosome, vacuole, and tail morphology learned from the MHSMA dataset.

Normal Sperm	90%
Abnormal Sperm	6%
Non-Sperm	3%

Note: This model is trained on the **MHSMA** dataset and is intended for research and educational purposes only. It does not replace professional medical diagnosis.

Running inference on server...

Figure 1. Normal Sperm Output



Sperm Morphology Classifier

Upload a microscopic sperm image to automatically classify it as normal, abnormal, or non-sperm using a trained CNN model.

Click to browse or drag an image
Supported: .png, .jpg, .jpeg • Recommended: single sperm cell microscopy image

Tip: Use grayscale or high-contrast images for best morphology detection.

Selected Image
Non-Sperm (10).bmp
60.1 KB

Model Output
CNN prediction based on your uploaded image

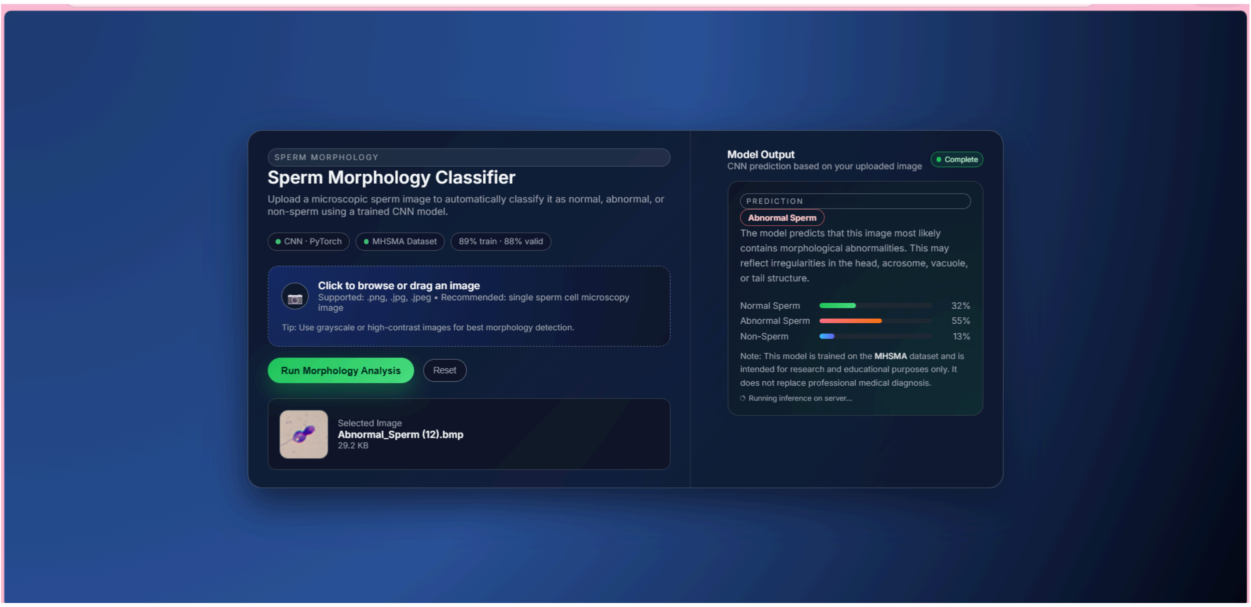
The model predicts that this image is more likely to be non-sperm material (e.g., debris or artifacts) rather than a valid sperm cell.

Normal Sperm	5%
Abnormal Sperm	4%
Non-Sperm	90%

Note: This model is trained on the **MHSMA** dataset and is intended for research and educational purposes only. It does not replace professional medical diagnosis.

Running inference on server...

Figure 2. Non-Sperm Output



SPERM MORPHOLOGY

Sperm Morphology Classifier

Upload a microscopic sperm image to automatically classify it as normal, abnormal, or non-sperm using a trained CNN model.

● CNN · PyTorch ● MHSMA Dataset 89% train · 88% valid

Click to browse or drag an image
Supported: .png, .jpg, .jpeg • Recommended: single sperm cell microscopy image

Tip: Use grayscale or high-contrast images for best morphology detection.

Selected Image
Abnormal_Sperm (12).bmp
29.2 KB

Model Output
CNN prediction based on your uploaded image

● Complete

PREDICTION

Abnormal Sperm

The model predicts that this image most likely contains morphological abnormalities. This may reflect irregularities in the head, acrosome, vacuole, or tail structure.

Normal Sperm	32%
Abnormal Sperm	55%
Non-Sperm	13%

Note: This model is trained on the **MHSMA** dataset and is intended for research and educational purposes only. It does not replace professional medical diagnosis.

Running inference on server...

Figure 3. Abnormal Sperm

XV. References

Finelli, R., Leisegang, K., Tumallapalli, S., Henkel, R., & Agarwal, A. (2021). The validity and reliability of computer-aided semen analyzers in performing semen analysis: A systematic review. *Translational Andrology and Urology*, *10*(7), 3069–3079. https://doi.org/10.21037/tau-21-276

Belala, R., Bourahmoune, D., & Mimoune, N. (2024). The use of computer assisted sperm analysis (CASA) in domestic animal reproduction: A review. *Kafkas Universitesi Veteriner Fakultesi Dergisi*, *30*(6), 741–751. https://doi.org/10.9775/kvfd.2024.32819

Valverde, A., Barquero, V., & Soler, C. (2020). The application of computer-assisted semen analysis (CASA) technology to optimise semen evaluation: A review. *Journal of Animal and Feed Sciences*, *27*(3), 190–198. https://doi.org/10.22358/jafs/127691/2020