Interpretable Sperm Morphology Classification via Attention-Guided Deep Learning

Zahra Asghari Varzaneh [1] **Reza Khoshkangini** [1] **Thomas Ebner** [2] **Lars Johansson** [3]

Abstract

Male infertility is a major cause of couple infertility, often linked to abnormal sperm morphology. While deep learning models offer automated analysis, most lack interpretability, limiting their clinical adoption. This study proposes an attention-guided deep learning framework for sperm morphology classification. We combine a pretrained EfficientNet-B0 with a Convolutional Block Attention Module (CBAM) to focus on key areas of the sperm head, improving both accuracy and interpretability. Evaluated on the SMIDS and HuSHem public datasets, our model achieves accuracies of 90.2% and 93.9% (macro F1-scores of 0.913 and 0.948), outperforming SimpleCNN and standard EfficientNet-B0. Furthermore, we use Grad-CAM++ visualizations to highlight features influencing the model's decisions. The results demonstrate that this accurate and transparent framework is a practical tool for automated sperm analysis in fertility clinics.

1. Introduction

Infertility is a global health issue affecting nearly 15% of couples, with male factors contributing to about half of these cases. Around 30% of male infertility cases relate directly to poor sperm quality and abnormal parameters (Agarwal et al., 2015). Various biological, environmental, and lifestyle factors can negatively impact sperm count, motility, and morphology, reducing pregnancy chances (Bonde, 1993; Durairajanayagam, 2018). Consequently, semen analysis is the primary test for evaluating male fertility, where sperm morphology assessment plays a critical role in predicting reproductive outcomes (Patel et al., 2018). Usually, this process is performed manually by laboratory specialists using light microscopy based on standard clinical guidelines (Kruger et al., 1988).

However, manual analysis is time-consuming, subjective, and highly dependent on the observer's experience. Studies show that observer agreement can be as low as 60–70% (Tomlinson et al., 2010). This lack of consistency, especially during high daily workloads, increases human error and limits diagnostic reliability. Therefore, objective and reproducible automated methods for sperm morphology analysis are highly needed.

Recent advances in artificial intelligence and deep learning (DL) have shown strong performance in medical image processing (Varzaneh et al., 2025b; Zamani et al., 2025; Varzaneh et al., 2025a). Recently, these techniques have been applied to sperm morphology classification. For instance, Liu et al. (Liu et al., 2021) used a modified AlexNet with transfer learning to classify human sperm morphology and reduce computational costs. Similarly, Cansiz et al. (Cansiz et al., 2025) introduced a generative adversarial framework to improve classification accuracy. Despite such progress, most DL models operate as "black boxes," providing predictions without explaining the underlying decision process. This lack of transparency limits their adoption in clinical practice (Sendak et al., 2020). To resolve this issue, explainable artificial intelligence (XAI) methods like Grad-CAM++ (Chattopadhay et al., 2018) have been developed. These techniques generate visual heatmaps to highlight key regions, helping clinicians verify model decisions.

In this study, we propose an attention-guided deep learning framework for sperm morphology classification. Our model combines EfficientNet-B0 (Tan & Le, 2019) with the Convolutional Block Attention Module (CBAM) (Woo et al., 2018) to focus on the most important regions of the sperm head. We evaluate our method on the SMIDS and HuSHem datasets, comparing it against a custom CNN and standard EfficientNet-B0. Additionally, Grad-CAM++ is used for visual explanations. The main contributions of this work are:

- Proposing an EfficientNet-B0 + CBAM framework for accurate and interpretable sperm morphology classification.
- Integrating a freeze-then-unfreeze training strategy to enhance performance on small datasets like HuSHem.

[1]Department of Computer Science and Media Technology, Malmö University, Malmö, Sweden [2]Kepler Universitätsklinikum, Linz, Austria [3]NewLifeAid-Global AB, Sweden. Correspondence to: Zahra Asghari Varzaneh <zahra.asghari-varzaneh@mau.se>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

• Generating class-specific visual explanations using Grad-CAM++ to improve clinical trust.

2. Methodology

We propose an attention-guided deep learning framework for automatic sperm morphology classification, combining EfficientNet-B0 for feature extraction with CBAM to focus on diagnostically important regions of the sperm head. To ensure clinical transparency, Grad-CAM++ is used for post-hoc visualization. As shown in Figure 1, the overall pipeline consists of four main stages: (1) data preprocessing and augmentation, (2) feature extraction via EfficientNet-B0 + CBAM, (3) training using a freeze-then-unfreeze strategy for small datasets, and (4) visual explanation via Grad-CAM++.

2.1. Preprocessing and Augmentation

Images are resized to 224×224 pixels and normalized using ImageNet statistics. Training augmentations include random flipping, rotation, color jittering, and affine transformations. For the small HuSHem dataset, MixUp augmentation (Eq. 1) and label smoothing are applied as regularization strategies:

$$\tilde{x} = \lambda x_i + (1 - \lambda) x_j \tag{1}$$

where λ is sampled from a Beta distribution.

2.2. EfficientNet-B0 Backbone

We use a compact pretrained EfficientNet-B0 (Tan & Le, 2019) initialized with ImageNet weights to leverage its compound scaling method. Given an input image X, the backbone extracts deep feature representations $F \in \mathbb{R}^{C \times H \times W}$ via:

$$F = f_\theta(X) \tag{2}$$

The resulting 1,280-channel feature map is then passed to the CBAM module.

2.3. CBAM Attention Module

To prioritize key features, CBAM (Woo et al., 2018) sequentially applies channel and spatial attention. The channel-attended feature map F' is computed as:

$$F' = M_c(F) \otimes F \tag{3}$$

Next, the spatial attention module refines these features focusing on critical regions:

$$F'' = M_s(F') \otimes F' \tag{4}$$

where M_c and M_s represent channel and spatial attention maps, and $\otimes$ is element-wise multiplication.

2.4. Freeze-then-Unfreeze Training

To prevent overfitting on small data, a two-stage strategy is used. First, the backbone and CBAM layers are frozen while only the classification head is trained. Second, all parameters are fine-tuned jointly with a lower learning rate using cross-entropy loss:

$$L_{CE} = -\sum_{i=1}^{K} y_i \log(\hat{y}_i) \tag{5}$$

where K is the class count, y_i is the true label, and $\hat{y}_i$ is the predicted probability. Optimization is done via AdamW with cosine annealing.

2.5. Grad-CAM++ for Explainability

Grad-CAM++ (Chattopadhay et al., 2018) generates visual heatmaps to highlight predictive regions. Class-specific importance weights are calculated from the final convolutional feature maps using:

$$L^c_{GradCAM++} = \mathrm{ReLU}\left(\sum_k w^c_k A^k\right) \tag{6}$$

where A^k is the k-th feature map and w^c_k is its weight for class c.

3. Experimental Setup and Datasets

We evaluate the framework on two public datasets. **SMIDS** (Ilhan & Aydın, 2018) contains 3,000 microscopic images categorized into three classes: Normal, Abnormal, and non-sperm. **HuSHem** (Shaker et al., 2017) contains 216 expert-verified sperm head images across four classes: Normal, Tapered, Pyriform, and Amorphous.
Experiments are implemented in PyTorch 2.6 on an NVIDIA GPU with a batch size of 32 for 100 epochs. Data is split into 70% training, 15% validation, and 15% testing using a fixed random seed. We compare three models: (1) a baseline SimpleCNN trained from scratch, (2) standard pretrained EfficientNet-B0, and (3) our proposed EfficientNet-B0 + CBAM model.

4. Results and Discussion

Table 1 summarizes accuracy and macro F1-score on the test sets for both datasets. The proposed method consistently achieves the best performance. On SMIDS, it reaches 90.2% accuracy and F1 = 0.913, outperforming both SimpleCNN and EfficientNet-B0 baseline while also providing interpretability through CBAM attention and Grad-CAM++ visualizations. On HuSHem, the proposed method achieves 93.9% accuracy and F1 = 0.948, compared to 63.6% for EfficientNet-B0 and 72.7% for SimpleCNN.

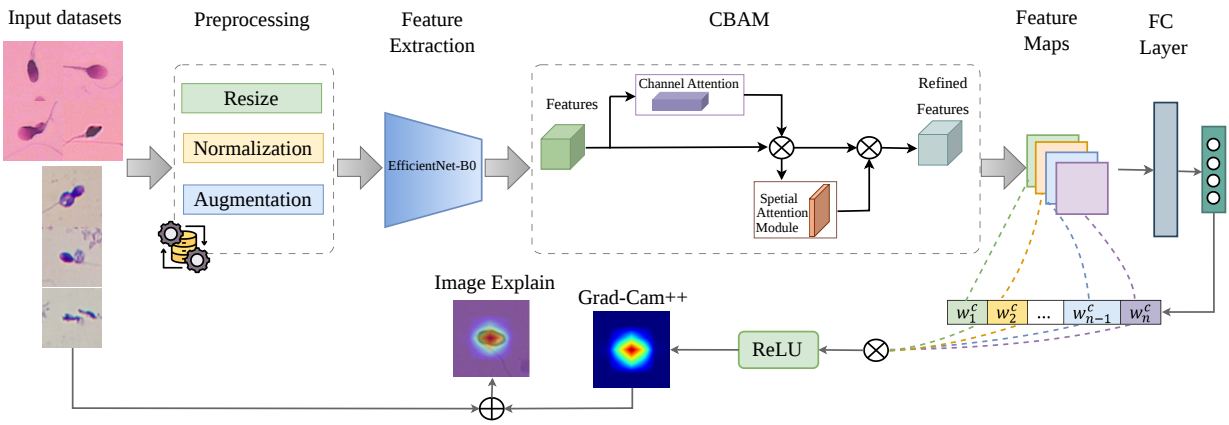

Figure 1. Overall architecture of the proposed framework.

This substantial performance gap on HuSHem demonstrates that the CBAM attention mechanism is especially effective for small datasets, as it forces the model to focus on morphologically relevant sperm head regions rather than background noise. The results also confirm the effectiveness of the freeze-then-unfreeze training strategy for medical image classification. To provide a more detailed analysis, Table 2 reports the per-class performance. On SMIDS, the proposed model achieves the highest performance for the Non-Sperm class (F1=0.97), which is clinically important because misclassifying debris as sperm can affect sperm count estimation. The Normal and Abnormal classes show balanced performance, indicating no strong bias toward a specific class.

On HuSHem, the proposed framework achieves strong discrimination across all morphology categories, with near-perfect performance on the Normal class and consistently high scores on Tapered, Pyriform, and Amorphous classes. This confirms that the attention-guided framework effectively captures subtle morphological differences in sperm head shapes. Figure 2 illustrates the training and validation curves of the proposed model on both datasets. On SMIDS, the model converges rapidly and remains stable, indicating strong generalization. On HuSHem, the effect of the freeze-then-unfreeze strategy is clearly visible: validation accuracy increases gradually during the frozen stage and improves significantly after unfreezing the backbone, stabilizing above 90%. Figure 3 shows ROC curves for all classes under a one-vs-rest setting. The proposed model achieves high AUC values on both datasets, with mean AUC of 0.965 for SMIDS and 0.991 for HuSHem, indicating excellent separability even under limited-data conditions.

Table 1. Classification results on SMIDS and HuSHem test sets. Best results in bold.

Model	SMIDS		HuSHem	
	Accuracy	F1-score	Accuracy	F1-score
SimpleCNN	82.67%	0.830	72.73%	0.695
EfficientNet-B0	88.00%	0.883	63.64%	0.684
Proposed	**90.21%**	**0.913**	**93.94%**	**0.948**

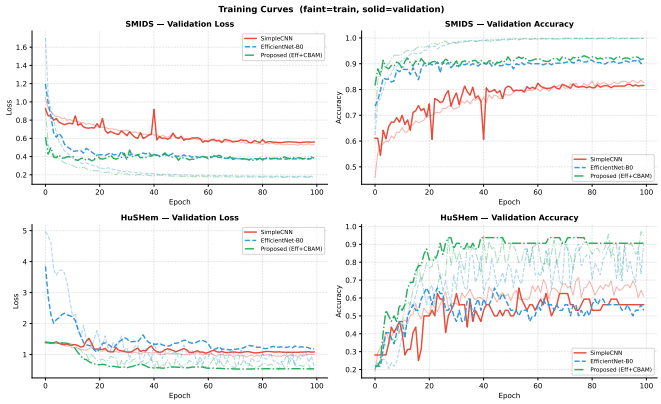

Figure 2. Training and validation curves for SMIDS and HuSHem datasets. The proposed model shows stable convergence and strong generalization. The two-phase behavior on HuSHem reflects the freeze-then-unfreeze strategy.

4.1. Explainable AI and Grad-CAM++ Visualizations

To improve interpretability, Grad-CAM++ visualizations are generated from the final convolutional layer of the proposed

Table 2. Per-class performance comparison (Precision | Recall | F1-score) for all models on SMIDS and HuSHem datasets.

Dataset	Class	SimpleCNN	EfficientNet-B0	Proposed
SMIDS	Normal	0.78 | 0.85 | 0.81	0.84 | 0.88 | 0.86	0.88 | 0.88 | 0.88
	Abnormal	0.81 | 0.74 | 0.77	0.86 | 0.82 | 0.84	0.87 | 0.87 | 0.87
	Non-Sperm	0.91 | 0.91 | 0.91	0.95 | 0.95 | 0.95	0.97 | 0.96 | 0.97
HuSHem	Normal	0.40 | 1.00 | 0.57	1.00 | 1.00 | 1.00	1.00 | 1.00 | 1.00
	Tapered	0.75 | 0.50 | 0.60	0.50 | 0.83 | 0.62	0.86 | 1.00 | 0.92
	Pyriform	1.00 | 0.58 | 0.74	0.67 | 0.67 | 0.67	0.92 | 0.92 | 0.92
	Amorphous	0.83 | 0.91 | 0.87	0.57 | 0.36 | 0.44	1.00 | 0.91 | 0.95

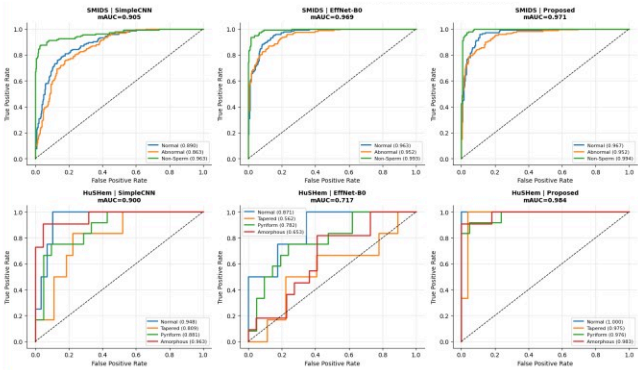

Figure 3. ROC curves for all models on SMIDS and HuSHem datasets using a one-vs-rest evaluation scheme.

network. Figure 4 presents representative results from both datasets.

The results show that the model consistently focuses on clinically relevant sperm head regions rather than background artifacts. In Normal samples, attention is concentrated on smooth oval head structures. In Abnormal, Pyriform, and Amorphous classes, attention shifts toward irregular boundaries and deformed regions. In Tapered sperm cells, the highest activation is localized at the elongated head tip, confirming that the model captures class-specific morphological patterns.

These observations confirm that CBAM improves spatial attention and enhances interpretability, making the decision process more transparent and clinically meaningful.

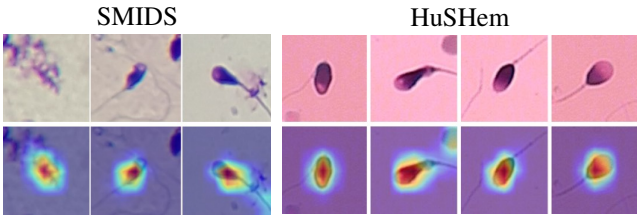

Figure 4. Grad-CAM++ visualizations for SMIDS and HuSHem datasets. Each dataset is organized into two rows: original images (top) and explanation heatmaps (bottom).

4.2. Ablation Study

The ablation study tracks how each component contributes to the overall success. Replacing SimpleCNN with a pretrained EfficientNet-B0 backbone improves SMIDS accuracy by 5.33%, but reduces performance on HuSHem by 9.09%. This suggests that directly fine-tuning a large pretrained model on only 150 training images leads to overfitting on background variations and staining artifacts.

Adding the CBAM attention module addresses this issue, improving HuSHem performance by 16.16%, while slightly increasing SMIDS (+0.22%). CBAM helps the model focus on the most relevant spatial locations and reweights feature channels toward the sperm head.

The freeze-then-unfreeze training strategy adds another 9.60% gain on HuSHem. Starting with a frozen backbone allows the model to learn a stable initialization before full fine-tuning. Additionally, MixUp regularization contributes a 4.54% improvement by smoothing decision boundaries, which is helpful in distinguishing between subtle morphology classes like Tapered and Pyriform.

Ultimately, the full model achieves a total improvement of 21.21% on HuSHem and 7.54% on SMIDS compared to the baseline. One limitation of this work is the small size of the HuSHem test set (33 samples), which introduces some uncertainty. Future studies should validate the method on larger, multi-centric datasets.

Table 3. Ablation study: cumulative effect of adding each component to SimpleCNN baseline. $\Delta\nabla$ values show change relative to the previous row. Green = improvement, red = regression.

Configuration	SMIDS			HuSHem		
	Accuracy	F1-score	$\Delta\nabla$	Accuracy	F1-score	$\Delta\nabla$
SimpleCNN	82.67%	0.830	—	72.73%	0.695	—
+ EfficientNet-B0	88.00%	0.883	-5.33%	63.64%	0.684	-9.09%
+ CBAM module	88.22%	0.885	+0.22%	79.80%	0.776	+16.16%
+ Freeze-unfreeze	88.16%	0.883	-0.06%	89.40%	0.882	+9.60%
Full model	90.21%	0.913	+2.05%	93.94%	0.948	+4.54%
Total gain vs. Baseline	+7.54%	+0.083		+21.21%	+0.253	

5. Conclusion

Manual diagnosis of male infertility remains subjective and highly variable. This study addressed this limitation by proposing an automated framework that balances classification accuracy with clinical transparency. By combining a pretrained EfficientNet-B0 backbone with CBAM attention and a two-stage training strategy, our approach demonstrates high effectiveness, particularly when labeled medical data is scarce. The results on the small HuSHem dataset highlight a key lesson: while standard pretrained networks can overfit and perform worse than simple baselines on limited data, integrating spatial/channel attention and targeted fine-tuning completely reverses this regression. Furthermore, Grad-CAM++ visualizations confirm that the model's focus aligns with manual clinical criteria. Such interpretability is crucial for moving automated tools from research environments into real fertility laboratories, making transparency a core requirement rather than an afterthought.

Acknowledgements

This work was supported by Vinnova, the Swedish Governmental Agency for Innovation Systems [Grant No. 2024-01462]. The funding source had no role in the design, execution, or publication of the study.

References

Agarwal, A., Mulgund, A., Hamada, A., and Chyatte, M. R. A unique view on male infertility around the globe. *Reproductive Biology and Endocrinology*, 13(1):37, 2015.

Bonde, J. P. The risk of male subfecundity attributable to welding of metals. studies of semen quality, infertility, fertility, adverse pregnancy outcome and childhood malignancy. *International Journal of Andrology*, 16:1–29, 1993.

Cansiz, B., Ilhan, H. O., and Serbes, G. Loss-based ensemble generative adversarial network model for enhancing the sperm morphology classification. *Advanced Intelligent Systems*, pp. e202500441, 2025.

Chattopadhay, A., Sarkar, A., Howlader, P., and Balasubramanian, V. N. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 839–847. IEEE, 2018.

Durairajanayagam, D. Lifestyle causes of male infertility. *Arab Journal of Urology*, 16(1):10–20, 2018.

Ilhan, H. O. and Aydın, N. A novel data acquisition and analyzing approach to spermiogram tests. *Biomedical Signal Processing and Control*, 41:129–139, 2018.

Kruger, T. F., Acosta, A. A., Simmons, K. F., Swanson, R. J., Matta, J. F., and Oehninger, S. Predictive value of abnormal sperm morphology in in vitro fertilization. *Fertility and Sterility*, 49(1):112–117, 1988.

Liu, R., Wang, M. M., Wang, M., Yin, J. Q., Yuan, Y. X., and Liu, J. Automatic microscopy analysis with transfer learning for classification of human sperm. *Applied Sciences-Basel*, 11(12), 2021.

Patel, A. S., Leong, J. Y., and Ramasamy, R. Prediction of male infertility by the world health organization laboratory manual for assessment of semen analysis: a systematic review. *Arab Journal of Urology*, 16(1):96–102, 2018.

Sendak, M., Elish, M. C., Gao, M., Futoma, J., Ratliff, W., Nichols, M., Bedoya, A., Balu, S., and O'Brien, C. The human body is a black box: supporting clinical decision-making with deep learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 99–109, 2020.

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

Tan, M. and Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pp. 6105–6114. PMLR, 2019.

Tomlinson, M. J., Pooley, K., Simpson, T., Newton, T., Hopkisson, J., Jayaprakasan, K., Jayaprakasan, R., Naeem, A., and Pridmore, T. Validation of a novel computer-assisted sperm analysis (casa) system using multitarget-tracking algorithms. *Fertility and Sterility*, 93(6):1911–1920, 2010.

Varzaneh, Z. A., Mousavi, S. M., Khoshkangini, R., and Khaliji, S. M. M. An ensemble model based on transfer learning for the early detection of alzheimer's disease. *Scientific Reports*, 15(1):34634, 2025a.

Varzaneh, Z. A., Wölner-Hanssen, N., and Khoshkangini, R. A lightweight transformer approach for predicting blastocyst formation on limited embryo images. In *2025 International Conference on Visual Communications and Image Processing (VCIP)*, pp. 1–5. IEEE, 2025b.

Woo, S., Park, J., Lee, J.-Y., and Kweon, I. S. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 3–19, 2018.

Zamani, H., Varzaneh, Z. A., and Nadimi-Shahraki, M. H. Deep convolutional neural networks in neurological disorders diagnosis: Comprehensive review of cutting-edge architectures, challenges, and future directions. *Archives of Computational Methods in Engineering*, pp. 1–83, 2025.