A Convolutional Neural Network Model with Feature Fusion for Sperm Morphology Classification

Firman Tempola
Department of Informatics, Khairun University, North Maluku, Indonesia | Department of Computer Science and Electronics, Gadjah Mada University, Yogyakarta, Indonesia
firman.tempola@unkhair.ac.id (corresponding author)

Ardiansyah
Department of Computer Science, Universitas Lampung, Lampung, Indonesia | Department of Computer Science and Electronics, Gadjah Mada University, Yogyakarta, Indonesia
ardiansyah@fmipa.unila.ac.id

Munazat Salmin
Department of Informatics, Khairun University, North Maluku, Indonesia | Department of Computer Science and Electronics, Gadjah Mada University, Yogyakarta, Indonesia
munazat@unkhair.ac.id

Leonardo Petra Refialy
Department of Informatics, Universitas Kristen Indonesia Maluku, Maluku, Indonesia | Department of Computer Science and Electronics, Gadjah Mada University, Yogyakarta, Indonesia
leonardopetrarefialy@mail.ugm.ac.id

Received: 7 October 2025 | Revised: 6 November 2025 and 26 November 2025 | Accepted: 27 November 2025

 DOI: https://doi.org/10.48084/etasr.15363

ABSTRACT

Sperm morphology analysis is a key parameter in male fertility evaluation, but the manual process is subjective and time-consuming. Therefore, an automated deep learning classification system offers a potential solution. This study evaluates and compares MobileNetV2 and EfficientNetB0, individually and in combination via feature fusion, for sperm morphology image classification using the SMID dataset of 3000 images. MobileNetV2 and EfficientNetB0 models achieved accuracies of 62.75% and 33.06%, respectively, after 50 epochs, while the feature fusion model reached 85.04% in only 5 epochs. Thus, combining features from both architectures yields superior accuracy and efficiency for automated sperm analysis.

Keywords-sperm morphology classification; deep learning; convolutional neural network; feature fusion; MobileNetV2; EfficientNetB0

I. INTRODUCTION

Infertility is a significant global health issue, with approximately 15% of reproductive-age couples experiencing difficulties in conceiving. The male factor contributes to nearly 50% of these infertility cases. A critical parameter in the evaluation of male infertility is semen analysis, particularly the assessment of sperm morphology. Sperm morphology refers to the shape and structure of the sperm, including its head, midpiece, and tail. Abnormalities in any of these components can impair the sperm's ability to fertilize an ovum. Therefore, assessing and understanding sperm morphology is vital in diagnosing and addressing male infertility.

Conventionally, sperm morphology assessment is performed manually under a microscope by embryologists or laboratory analysts. However, this method is highly subjective and susceptible to both inter-observer and intra-observer variability. Furthermore, the process is time-consuming, requires a high level of skill, and demands intense focus, rendering it inefficient and impractical for large-scale applications [1]. Previous research on male fertility classification has explored various approaches. Some studies

have utilized non-image data, such as patient lifestyle factors [2-4], while others have focused on sperm motility analysis [5]. Other works have concentrated on sperm segmentation to improve image quality for subsequent classification [6, 7].

Sperm quality classification from morphological images can be performed using both machine learning and deep learning models. Machine learning models, as demonstrated in [8], typically require a preliminary feature extraction stage. In contrast, deep learning models can learn features automatically but often demand substantial computational resources and large datasets [9, 10]. Dataset imbalances in sperm quality classification have been also addressed [11].

Sperm analysis was also conducted in [12], where focus was placed on sperm movement (motility) analysis using YOLOv5, a deep learning object detection algorithm, with a dataset sourced from VISEM-Tracking (a collection of videos with labeled bounding box annotations for tracking). However, this study did not include a detailed discussion of sperm morphology. Authors in [13] used SiD V1.0 software to analyze sperm kinematics, such as Straight-Line Velocity (VSL), Linearity (LIN), and Hyperactivation Motility Pattern (HMP), concentrating on sperm selection for Intracytoplasmic Sperm Injection (ICSI) based on movement. However, they did not use deep learning to identify differences in sperm shape (morphological classification). Authors in [14] proposed SFCNet, a specialized convolutional neural network for sperm segmentation and tracking, achieving high performance in distinguishing sperm in images and analyzing motility. This approach did not employ a fusion model or feature fusion techniques for classifying sperm morphology. Meanwhile, authors in [15] applied machine learning techniques to select embryos based on sperm analysis.

The present study proposes a multi-fusion convolutional neural network. While multi-fusion in deep learning has been previously explored [16], it involved fusing outputs from a single model architecture trained with different image and batch sizes, followed by a voting mechanism, which necessitates significant computational overhead. The method proposed in the current study differs by combining two distinct architectures: EfficientNetB0 and MobileNetV2. These models offer complementary strengths; EfficientNetB0 is recognized for its computational efficiency, while MobileNetV2 is a lightweight architecture. Combining their features accelerates the classification of sperm abnormalities without requiring additional resources, thereby addressing a key limitation of many deep learning models. The combination of MobileNetV2 and EfficientNetB0 has been previously implemented in [17] for the classification of mango leaf diseases. However, the classification was performed using a machine learning model and included a feature selection stage. This contrasts with the proposed approach, which uses a CNN model without feature selection.

II. MATERIALS AND METHODS

A. Dataset

The dataset utilized in this study is the Sperm Morphology Image Dataset (SMID) [18]. It comprises three categories: non_sperm, normal_sperm, and abnormal_sperm. The SMID dataset contains 3000 images, with 1000 per category. Data acquisition uses several hardware items: smartphones, microscopes, and stabilizers. The resulting data, in the form of video, are then extracted into several frames with an image resolution of 1920 × 1080 pixels. The detailed image acquisition process is related to the SMID dataset [19]. The types of datasets used from each category are shown in Figure 1.

Fig. 1. Sample images from the dataset representing (from left to right) normal_sperm, abnormal_sperm, and non_sperm object.

B. Proposed Method

The proposed model for classifying sperm morphology abnormalities introduces an innovative approach that fuses features from the MobileNetV2 and EfficientNetB0 architectures. The system processes an input image of 224 × 224 × 3 size. The parallel feature extraction process is illustrated in Figure 2. As depicted in the architecture, an input image is fed into both models simultaneously. Within each model, the image passes through several convolutional layers, producing a 7 × 7 × 1280 feature map. This indicates that 1280 distinct filters generate features. Subsequently, a Global Average Pooling (GAP) layer reduces each 2D feature map to a single average value using (1), where W and H are the width and height of the feature map, respectively:

$$GAP = \frac{1}{W\,x\,H} \sum_{i=1}^{W} \sum_{j=1}^{H} Feature\ Map\ (i,j) \qquad (1)$$

This GAP operation produces a feature vector of size 1280 for MobileNetV2 and 1280 for EfficientNetB0. The feature vectors from both models are then fused (concatenated) to create a combined feature vector of size 2560. This fused representation is then passed to a classification head, and the model's performance is evaluated. The complete classification pipeline is displayed in Figure 2.

Based on the classification model portrayed in Figure 3, the initial stage of this experiment is to load the dataset, then apply both CNN architectures as extractors. The architectures of both models were modified several time, namely the last fully connected layer was removed, fine-tuning was not, and the feature map was converted to a 1D vector. Next, the features from both models were combined, and the CNN stages were applied. This experiment was also validated with k-fold cross-validation with an initialized k value of 5. So that each k gets a total of 600 data points. Thus, in the test, the training data are 2400 and the testing data are 600. Each evaluation with k-fold cross-validation calculates its metrics. The metrics used in this test are precision, recall, F1-score, and accuracy. If the experiment has been completed up to the 5th k, the average of each metric is calculated, as illustrated in Figure 3.

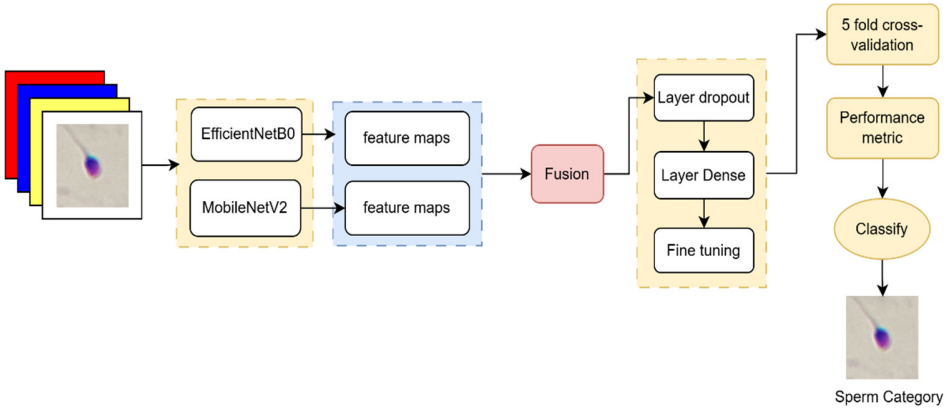

Fig. 2. The proposed model for sperm morphology classification.

C. Experimental Setup

All training and testing procedures were conducted on Google Colaboratory with 12.72 GB of RAM and GPU acceleration. The same hyperparameters were used for both the individual and fused models, as detailed in Table I.

TABLE I. HYPERPARAMETER CONFIGURATION

Parameter	Configuration
Image size	224 × 224
Batch size	32
Activation function	Softmax
Learning rate	0.001
Epoch size	5, 10, 20, 30, 50
Dropout	0.4
Dense	512
k-fold cross-validation	5

III. RESULTS AND DISCUSSION

A. Performance of Individual MobileNetV2 and EfficientNetB0 Models

To provide context for the proposed model, first, a preliminary experiment on sperm morphology classification was conducted using two baseline CNN models: MobileNetV2 and EfficientNetB0. The experiments were carried out over several epochs. Figure 3 presents the resulting graphs from these two baseline CNN experiments, serving as a comparison point for the subsequent feature fusion approach.

The experimental results for MobileNetV2 are shown in Figure 3 (a) for one of the five folds, trained for 50 epochs. The model achieved an average accuracy of 62.75% for sperm morphology classification. Several factors contributed to this relatively low accuracy. First, the dataset contains only 3001 images. This is contrary to MobileNetV2's need for a large dataset for effective training. Second, the suboptimal quality of the medical images made it difficult for the model to recognize each sperm morphology class. Third, the subtle distinction between abnormal and normal classes also reduced the model's performance.

In contrast, implementing EfficientNetB0 with identical parameters yielded a mean accuracy of only 33.06%. This result is significantly lower than that achieved by MobileNetV2. Although EfficientNetB0 is efficient, it is more complex and has greater learning capacity than MobileNetV2. When a complex model is trained on a small dataset, it often exhibits a strong tendency to memorize the training data rather than generalize to the underlying patterns—a phenomenon known as overfitting. Consequently, its performance is deceptively high on data it has already seen (training data) but deteriorates substantially when evaluated on new, unseen data (validation/test data). This is evidenced by the experimental results displayed in Figure 3 (b), where the training performance was superior to the testing and validation results.

B. Analysis of Feature Fusion Model Results

The feature fusion process was tested over several epochs with a learning rate of 0.001, the Adam optimizer, and 5-fold cross-validation. The results obtained from each experiment varied, as shown in Figure 4.

The experimental results in Figure 4 show that, at a constant learning rate of 0.001, model accuracy peaked at 85.04% after 5 epochs. Subsequent increases in the number of epochs led to a consistent decrease in performance: 83.27% at 10 epochs, 83.17% at 20 epochs, and 81.87% at 30 epochs. This decline suggests that the model is beginning to memorize the training data rather than learning generalizable patterns, reducing its ability to perform well on new data.

The 5-epoch test yielded the highest accuracy. This run took 4 h, 23 min, and 15 s due to the increasing feature set and the limited computing resources. The confusion matrix is portrayed in Figure 5.

Based on the results of the 5th confusion matrix fold depicted in Figure 5, the precision value is 86.45%, the recall is 86.50%, the F1 Score is 86.40%, and the accuracy is 86.50%. The accuracies for each class are 78.02% for normal_sperm, 91.24% for non_sperm, and 84.83% for abnormal_sperm. The results of this fifth fold also show that the normal_sperm class is frequently misclassified as abnormal_sperm, while the abnormal_sperm class is likewise misclassified as normal_sperm. The reason is that these two classes are visually similar, so during classification, the system is unable to distinguish them properly. However, the non_sperm class is easier to recognize as non-sperm. The performance of each fold is presented in Table II.

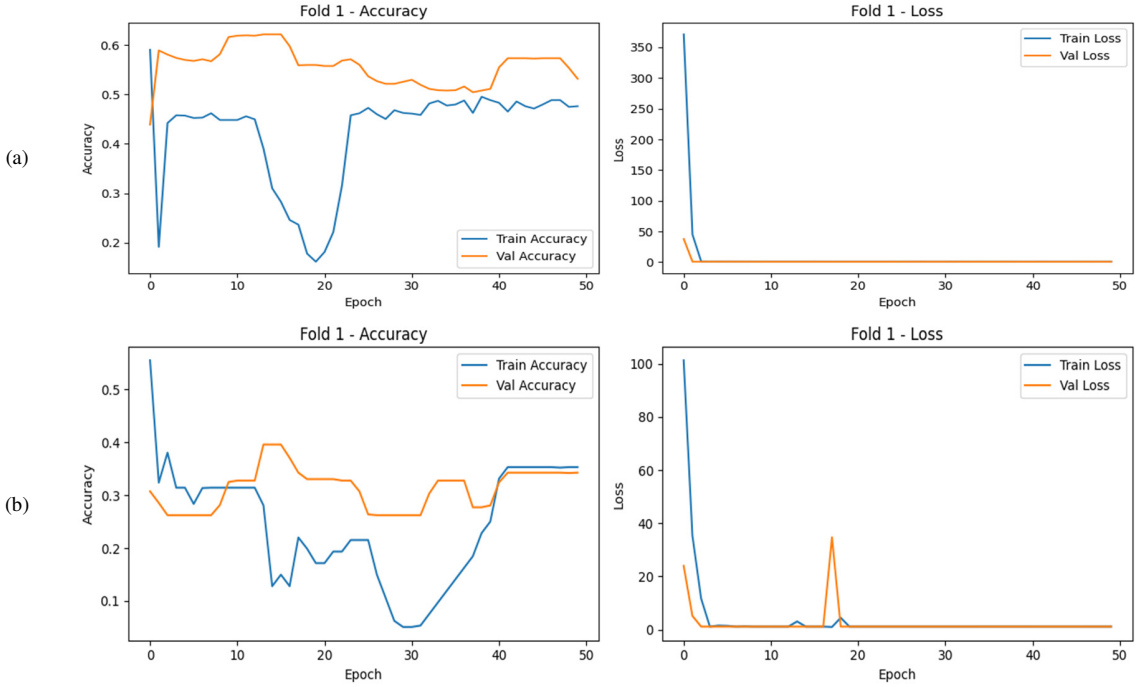

Fig. 3. Accuracy and loss curves for: (a) MobileNetV2 and (b) EfficientNetB0 over 50 epochs.

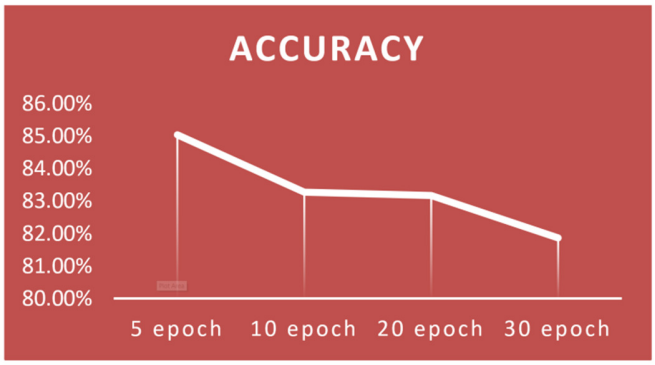

Fig. 4. Feature fusion experimental results.

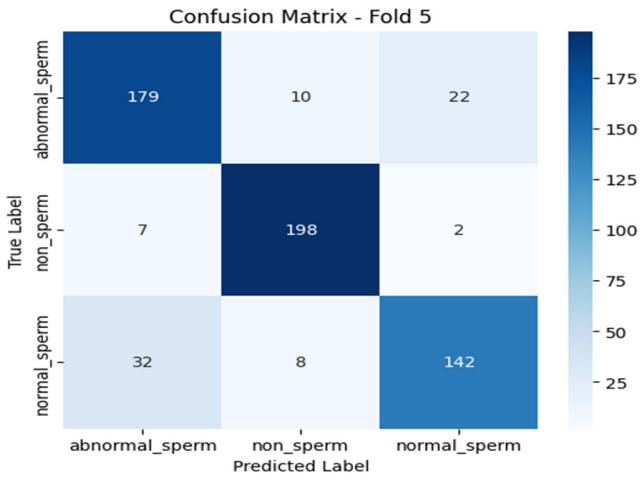

Fig. 5. Fusion feature experimental results for the 5th fold.

TABLE II. EXPERIMENTAL RESULTS OF FEATURE FUSION FOR ALL *k* FOLDS

Fold to	Accuracy	Precision	Recall	F1-score
1	84.19%	84.36%	84.19%	84.19%
2	83.67%	84.92%	83.67%	83.64%
3	84.67%	87.19%	84.67%	84.55%
4	86.17%	86.31%	86.17%	86.21%
5	86.50%	86.45%	86.50%	86.40%
Average	85.04%	85.85%	85.04%	85.00%

C. *Comparison of Feature Fusion with MobileNetV2 and EfficientNetB0*

Based on the test results, EfficientNetB0 yielded the lowest accuracy at 33.06%. Practically, this suggests that the model learned almost nothing, with its performance approaching that of a random guess. For a 3-class classification problem, random guessing yields an accuracy of approximately 33%, which closely matches this result, indicating that the EfficientNetB0 model underfits, as it fails to capture the underlying patterns and complexity of the sperm image data. This may also be attributed to suboptimal hyperparameters that prevented the model from converging, which is further supported by the fact that its accuracy remained static across experiments initialized with 10, 30, and 50 epochs.

In contrast, the 62.75% accuracy achieved by MobileNetV2 indicates that the model has learned some useful patterns, although this performance is still insufficient for practical application. The accuracy of MobileNetV2 could potentially be improved through hyperparameter tuning (e.g., learning rate, optimizer, batch size), increasing the number of epochs, or applying more aggressive data augmentation techniques. This potential for improvement is supported by the observation that

when initialized with 10 epochs, accuracy was 44.75% and subsequently increased as the number of epochs increased, implying that accuracy will likely continue to increase with more training epochs.

Finally, the best performance in sperm morphology classification was achieved by the feature fusion model, which obtained an accuracy of 85.04%. This surpasses the performance of the best individual model (MobileNetV2) by over 20%. The superior performance of the feature fusion model is due to its approach of combining the "knowledge" (i.e., features) extracted by both underlying models. Although EfficientNetB0 performed poorly on its own; however, this does not mean it learned nothing. It is highly likely that the two models captured different and complementary types of features. MobileNetV2 was possibly better at capturing general structural features and the global shape of the sperm head and tail. Meanwhile, EfficientNetB0, despite its overall failure, may have succeeded in extracting certain texture features or fine details that were missed by MobileNetV2.

D. Comparison with State-of-the-Art

The performance of the proposed model is compared with other works in this area, as shown in Table III. It is observed that the proposed method, which employs a feature fusion technique, demonstrates a clear and significant superiority over four previous studies in the task of sperm morphology classification.

TABLE III. COMPARISON WITH STATE-OF-THE-ART METHODS

Innovation	Input type	Accuracy
Classification of sperm quality based on lifestyle [2]	Text and numeric	61.20%
Deep learning [10]	Image	68.00%
Soft voting, multi-fusion CNN [16]	Image	66.45%
Augmentation and adversarial attack [20]	Image	76.36%
Dual Tree Complex Wavelet Transform [21]	Image	82.33%
Feature fusion (proposed method)	Image	85.04%

Authors in [2, 10] addressed different or narrower scopes. That is, authors in [2] focused on correlating sperm quality with lifestyle factors instead of strictly performing visual classification. Similarly, authors in [10] concentrated exclusively on the classification of the sperm head. In contrast, the proposed feature fusion method utilizes a CNN for holistic sperm morphology classification and has proven more effective for the overall visual classification task.

While authors in [16] also applied a fusion concept, they focused on the decision level (soft voting), where independent model predictions are combined only at the end. Conversely, the proposed method implements feature-level fusion, which integrates insights from models before classification. The resulting performance gap (83.54% compared to 66.45%) suggests that combining deep feature representations is far more effective than the decision-averaging approach used in [16].

Authors in [20] focused on enhancing data robustness through augmentation and adversarial attacks. While this approach achieved 76.36% accuracy, the proposed method demonstrates that innovation at the model architecture level (via feature fusion) has a substantially greater impact than merely strengthening the training data. This suggests that feature fusion is capable of extracting far richer and more informative feature representations from existing data. Authors in [21] applied the Wavelet Transform to sperm morphology classification using SVM as the classification model. While RBF achieved the highest accuracy, it was still lower than that of feature fusion.

IV. CONCLUSION

It was seen that combining features from MobileNetV2 and EfficientNetB0 architectures and then classifying with CNN significantly outperformed each individual model. The fusion model achieved an average accuracy of 85.04%. EfficientNetB0 alone had low performance at 33.06% under this study's parameters. Still, it managed to extract unique features that complemented those from MobileNetV2, with an accuracy of 62.75%. Combining features from both architectures created a more informative feature set, enabling the classifier to perform more effectively. Notably, the fusion model reached the highest accuracy with CNN and 5-fold classification in only 5 epochs. However, the entire experiment took more than 4 h, 23 min, and 15 s to complete. Therefore, future work should consider a feature selection stage to reduce irrelevant features and make the experimental process more efficient.

ACKNOWLEDGMENT

This research was sponsored by Lembaga Pengelola Dana Pendidikan (LPDP), Ministry of Finance, Republic of Indonesia. The authors extend their sincere gratitude for this support. And thanks to the Center for Higher Education Funding and Assessment (PPAPT), the Ministry of Higher Education, Science, and Technology of the Republic of Indonesia, and the Indonesia Endowment Fund for Education (LPDP) for funding his doctoral study through the Indonesian Education Scholarship (BPI).

REFERENCES

[1] G. M. Centola, "Semen Assessment,"*Urologic Clinics of North America*, vol. 41, no. 1, pp. 163–167, Feb. 2014, https://doi.org/10.1016/j.ucl.2013.08.007.

[2] A. Aykaç, C. Kaya, Ö. Çelik, M. E. Aydın, and M. Sungur, "The prediction of Semen Quality Based on Lifestyle Behaviours by the Machine Learning Based Models," *Reproductive Biology and Endocrinology*, vol. 22, no. 1, Aug. 2024, Art. no. 112, https://doi.org/10.1186/s12958-024-01268-w.

[3] H. E. Lyons, P. Gyawali, N. Mathews, P. Castleton, S. M. Mutuku, and N. O. McPherson, "The Influence of Lifestyle and Biological Factors on Semen Variability," *Journal of Assisted Reproduction and Genetics*, vol. 41, no. 4, pp. 1097–1109, Apr. 2024, https://doi.org/10.1007/s10815-024-03030-y.

[4] H.-H. Huang *et al.*, "Machine Learning Predictive Models for Evaluating Risk Factors Affecting Sperm Count: Predictions Based on Health Screening Indicators," *Journal of Clinical Medicine*, vol. 12, no. 3, Feb. 2023, Art. no. 1220, https://doi.org/10.3390/jcm12031220.

[5] H. O. Ilhan, M. Yuzkat, and N. Aydin, "Sperm Motility Analysis by using Recursive Kalman Filters with the Smartphone-Based Data Acquisition and Reporting Approach," *Expert Systems with Applications*, vol. 186, Dec. 2021, Art. no. 115774, https://doi.org/10.1016/j.eswa.2021.115774.

[6] R. Marín and V. Chang, "Impact of Transfer Learning for Human Sperm Segmentation using Deep Learning," *Computers in Biology and Medicine*, vol. 136, Sept. 2021, Art. no. 104687, https://doi.org/10.1016/j.compbiomed.2021.104687.

[7] E. Lewandowska, D. Węsierski, M. Mazur-Milecka, J. Liss, and A. Jezierska, "Ensembling Noisy Segmentation Masks of Blurred Sperm Images," *Computers in Biology and Medicine*, vol. 166, Nov. 2023, Art. no. 107520, https://doi.org/10.1016/j.compbiomed.2023.107520.

[8] A. Mehrjerd, T. Dehghani, M. Jajroudi, S. Eslami, H. Rezaei, and N. K. Ghaebi, "Ensemble Machine Learning Models for Sperm Quality Evaluation Concerning Success Rate of Clinical Pregnancy in Assisted Reproductive Techniques," *Scientific Reports*, vol. 14, no. 1, Oct. 2024, Art. no. 24283, https://doi.org/10.1038/s41598-024-73326-7.

[9] Y. Guo *et al.*, "Automated Deep Learning Model for Sperm Head Segmentation, Pose Correction, and Classification," *Applied Sciences*, vol. 14, no. 23, Dec. 2024, Art. no. 11303, https://doi.org/10.3390/app142311303.

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

[11] H. Jabbari and N. 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, Sept. 2023, https://doi.org/10.1007/s00521-023-08742-3.

[12] V. Thambawita *et al.*, "VISEM-Tracking, a Human Spermatozoa Tracking Dataset," *Scientific Data*, vol. 10, no. 1, May 2023, Art. no. 260, https://doi.org/10.1038/s41597-023-02173-4.

[13] G. Mendizabal-Ruiz *et al.*, "Computer Software (SiD) Assisted Real-Time Single Sperm Selection Associated with Fertilization and Blastocyst Formation," *Reproductive BioMedicine Online*, vol. 45, no. 4, pp. 703–711, Oct. 2022, https://doi.org/10.1016/j.rbmo.2022.03.036.

[14] W. Dai *et al.*, "Automated Non-Invasive Analysis of Motile Sperms using Sperm Feature-Correlated Network," *IEEE Transactions on Automation Science and Engineering*, vol. 22, pp. 3960–3970, 2025, https://doi.org/10.1109/TASE.2024.3404488.

[15] J. B. You, C. McCallum, Y. Wang, J. Riordon, R. Nosrati, and D. Sinton, "Machine Learning for Sperm Selection," *Nature Reviews Urology*, vol. 18, no. 7, pp. 387–403, July 2021, https://doi.org/10.1038/s41585-021-00465-1.

[16] M. Yüzkat, H. O. Ilhan, and N. Aydin, "Multi-model CNN Fusion for Sperm Morphology Analysis," *Computers in Biology and Medicine*, vol. 137, Oct. 2021, Art. no. 104790, https://doi.org/10.1016/j.compbiomed.2021.104790.

[17] S. Khandelwal, A. Raut, H. Vyawahare, D. Theng, and S. Dhande, "Optimizing Performance in Mango Plant Leaf Disease Classification through Advanced Machine Learning Techniques," *Engineering, Technology & Applied Science Research*, vol. 14, no. 6, pp. 18476–18480, Dec. 2024, https://doi.org/10.48084/etasr.8220.

[18] H. Ilhan, "Sperm Morphology Image Data Set (SMIDS)." Mendeley, Jan. 17, 2022, [Online]. Available: https://data.mendeley.com/datasets/6xvdhc9fyb/1.

[19] H. O. Ilhan and N. Aydin, "A Novel Data Acquisition and Analyzing Approach to Spermiogram Tests," *Biomedical Signal Processing and Control*, vol. 41, pp. 129–139, Mar. 2018, https://doi.org/10.1016/j.bspc.2017.11.009.

[20] A. Nabipour, M. J. Shams Nejati, Y. Boreshban, and S. A. Mirroshandel, "Less-Supervised Learning with Knowledge Distillation for Sperm Morphology Analysis," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 12, no. 1, Dec. 2024, Art. no. 2347978, https://doi.org/10.1080/21681163.2024.2347978.

[21] H. O. Ilhan, G. Serbes, and N. Aydin, "Dual Tree Complex Wavelet Transform Based Sperm Abnormality Classification," in *2018 41st International Conference on Telecommunications and Signal Processing*, Athens, Greece, July 2018, pp. 1–5, https://doi.org/10.1109/TSP.2018.8441431.