JIKO (Jurnal Informatika dan Komputer) Accredited KEMENDIKBUD RISTEK, No.105/E/KPT/2022
Vol. 7, No. 3, December 2024, pp. 196-203 p-ISSN: 2614-8897
DOI: 10.33387/jiko.v7i3.8938 e-ISSN: 2656-1948

SPERM ABNORMALITY CLASSIFICATION USING MULTI-PURPOSE IMAGE EMBEDDING AND CLASSICAL MACHINE LEARNING

Sigit Adinugroho[1*], Yuita Arum Sari[2], Wijaya Kurniawan [3], Achmad Arwan[4]

[1,2,3,4]Faculty of Computer Science, Brawijaya University, Malang, INDONESIA
Email: [*1]sigit.adinu@ub.ac.id, [2]yuita@ub.ac.id, [3]wjaykurnia@ub.ac.id [4]arwan@ub.ac.id

(Received: 8 October 2024, Revised: 10 November 2024, Accepted: 21 November 2024)

Abstract

Since sperm cells have big impact for human welfare in terms of reproduction, there are many studies have been done. In this case, we are attracted to enrich the method in determining the morphological properties of them using machine learning. Most study about it is done using 2-steps action that are feature extraction which is continued by classification. In our work, we aimed to lower the complexity by using image embedding as a general-purpose feature extractor that requires no training. For feature extraction using image, it is found that RGB has better performance compared to grayscale if we want to use Support Vector Machine (SVM). Meanwhile, when a comparation is done between SVM, random forest, Multi-Layer Perceptron (MLP), Naïve Bayes, and k-Nearest Neighbour (kNN) for classification process, MLP shows the best performance among them which is around 85%. Moreover, our proposed method has low complexity indicated by the training time around one and a quarter minute s for the most accurate method, compared to hours of training time in similar methods.

Keywords: *sperm cells morphology, machine learning, feature extraction, classification, SVM, MLP.*

**Corresponding Author: Sigit Adinugroho*

1 INTRODUCTION

The morphology of sperm cells significantly influences human fertility. Numerous research studies have been carried out to examine this phenomenon. The findings indicate that sperm abnormalities are more prevalent in men with fertility issues [1]. A similar investigation suggested that fertile men tend to have a higher proportion of normal sperm (sperm without defects) compared to infertile men [2]. Consequently, the morphological characteristics of sperm are crucial in predicting the success of assisted reproduction [3], [4]. Similarly, under in vivo conditions, a higher pregnancy rate is anticipated when the sperm exhibit good morphology. Furthermore, good sperm morphology is also associated with a shorter time to pregnancy in natural conception [5].

According to the criteria established by the World Health Organization (WHO) [6], sperm morphology evaluation involves the examination of spermatozoa by laboratory personnel using a light microscope set at a magnification of x1000. It is necessary for the technician to evaluate a minimum of 200 sperm cells and categorize them based on their form. Nonetheless, the subjective nature of human assessment introduces the possibility of inconsistency. Research conducted in Australia demonstrated variations in morphology assessments among different laboratories [7]. This same trend was observed in studies carried out in Italy [8], Belgium [9], and Spain [10].

The lack of consistency in manual sperm assessment has prompted the creation of an automated method known as Computer-Aided Semen Analysis (CASA). A typical CASA system is comprised of three main elements: a camera, microscope, and image processing system [11]. Currently, CASA systems are widely utilized in various laboratories globally. CASA systems offer advantages such as decreased subjectivity and minimized human error [12], as well as the capacity to analyze larger sample sizes, leading to shorter analysis times and enhanced productivity [13].

Numerous approaches have been developed for the estimation of morphological characteristics of sperm cells. Yüzkat et al. [14] combined the outcomes of six Convolutional Neural Network (CNN) models using two fusion techniques. A comparable ensemble method was utilized by merging four CNN models (VGG16, VGG19, ResNet34, and DenseNet-161) with

a meta classifier [15]. Yang et al. [16] utilized the BlendMask framework to identify individual cells, followed by segmentation of a cell into head, midpiece, and principal piece components using SegNet. Subsequently, a classification process was carried out by EfficientNet based on the components of the sperm. Likewise, U-Net was employed for the segmentation of a spermatozoon into head, neck, and tail regions [17], although no classification was performed for the estimation of morphology. A VGG-like network was proposed for the purpose of morphology classification [18], [19].

The majority of techniques utilized for morphology analysis rely on deep neural networks, particularly CNNs, for their ability to effectively extract features from images. The training process typically involves starting from scratch or employing transfer learning. In the former approach, a randomly initialized network is trained on a specialized dataset related to sperm [14], [17], whereas in the latter approach, a pre-trained network on a general dataset like ImageNet is fine-tuned using a dataset specific to the subject [15], [19].

Despite having strong performance in various applications, including sperm image classification, CNN require extensive training time due to their large number of parameters. A study indicated that the retraining process using transfer learning took approximately 30 minutes and 2 hours per fold for datasets containing 216 and 1132 images, respectively [19]. Overall, complex architectures like ensembled models typically necessitate longer training times.

In practical applications, it may not always be essential to train a CNN on a specific set of images that closely resemble the test set. Training on a diverse and extensive dataset is adequate to achieve comparable performance [20]. This principle has led to the advancement of image embedding, which entails utilizing a pretrained CNN on a large dataset like ImageNet as a feature extractor [21]. The approach involves modifying the final classifier network of a CNN to generate a fixed-length vector serving as a feature for subsequent processing. The utilization of image embedding along with various machine learning techniques has been investigated and proven effective in multiple domains, including image clustering [22], medical image classification [23], and remote sensing image classification [24].

The primary objective of this research is to categorize the morphology of a sperm cell by utilizing image embedding as a feature extraction technique to streamline the training process. Various machine learning algorithms are assessed to determine their effectiveness in classifying sperm cell morphology. Additionally, a comparative analysis is conducted to evaluate the efficacy of our proposed methodology in comparison to existing approaches.

2 RESEARCH METHOD

2.1 Overview of the methods

Our proposed approach comprised two primary components: a feature extractor and a classifier. The feature extractor component took in an RGB image and generated a feature vector of specific length. Conversely, the classifier component classified the feature vector into one of several categories indicating the state of a sperm cell: abnormal, normal, or non-sperm. The feature extractor was constructed utilizing image embedding techniques, whereas the classifier utilized traditional machine learning algorithms to expedite the training process. Overall flow of our proposed work is presented in Figure 1.

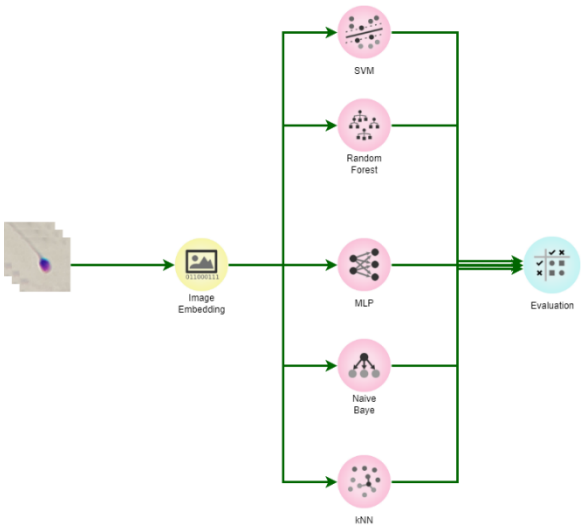


Figure 1. Visual overview of the proposed methods

2.2 Image Embedding

Image embedding refers to the process of converting the pixel-based representation of an image into a feature vector. The objective is to create a vector that effectively encapsulates the visual attributes and underlying semantics of the image. The fundamental approach employed in constructing image embedding involves the utilization of deep learning techniques, specifically Convolutional Neural Networks (CNNs) [25].

In a typical situation, an image embedding model is constructed through the training of a CNN model on classification tasks. Various CNN models, including VGG, ResNet, Inception, and SqueezeNet, can serve as feature extractors for image embedding. Once training is finished, the output layer is removed and activations from the preceding layer are obtained, normalized, and combined to create a feature vector. This resultant vector is robust against image transformations, changes in brightness, and noise [26]. Consequently, it serves as an optimal input for a range of machine learning techniques. Since the training is only conducted once and the network can be utilized to produce feature vectors for various computer vision tasks, the network can be seen as a multi-purpose image embedding [27].

SqueezeNet [28] is a convolutional neural network (CNN) model designed to achieve high accuracy while minimizing the number of parameters used. The model incorporates a unique fire module that combines 1x1 and 3x3 convolution filters to reduce parameter count. The architecture of SqueezeNet consists of a series of convolutional layers, 8 fire modules, and a final convolutional layer, allowing it to deliver performance similar to AlexNet but with significantly fewer parameters, specifically 50 times less.

A comparison of the performance and parameter count of various CNN models is presented in Table 1 [29]. It is evident from the comparison that SqueezeNet stands out for its notably lower parameter count while maintaining comparable performance to other models. This makes SqueezeNet an attractive choice for applications requiring image embedding.

Table 1. Comparison Of Parameters and Performance Of Several CNN Architectures

Model	Number of Parameters	Top-5 Accuracy on ImageNet
VGG 16	138M	89.8%
VGG 19	143M	89.8%
ResNet 18	11.7M	89.45%
ResNet 34	21.8M	91.4%
Inception V3	23.8M	93.9%
SqueezeNet	3.2M	88.20%

2.3 Classification Algorithms

Classification in machine learning involves the assignment of a specific category to a data point in a manner that is both mutually exclusive and exhaustive. There are numerous algorithms that can be utilized to classify feature vectors extracted from image embeddings into distinct categories related to sperm defects.

a. Support Vector Machine (SVM)

SVM tries to build hyperplane that maximally separates datapoint according to their class labels, thus it is usually referred as maximum margin classifier [30]. In binary classification, given training pairs $\{x_i, y_i\}_{i=1}^n$, where $x_i \in \mathbb{R}^N$ and $y_i \in \{-1,1\}$, SVM solves the Formula 1

$$min\left\{\frac{1}{2}\|w\|^2 + C\sum_i^n \xi i\right\} \quad (1)$$

SVM employs the kernel trick to convert the data from their initial space to a higher-dimensional space. In some cases, the data may not be easily separable by a linear method. Consequently, moving them to a higher-dimensional space could help in identifying a linearly separable hyperplane. Various well-known kernels include the linear, polynomial, sigmoid, Gaussian radial basis function, and randomized blocks analysis of variance [31].

The original SVM algorithm is limited to binary classification tasks involving only two classes. To address this limitation, modifications were made to the classic algorithm by incorporating methods like one-against-all, one-against-one, and multiclassification objective functions.

b. Random Forest (RF)

The random forest classifier is comprised of multiple decision tree classifiers that aggregate the output through a majority vote from each individual tree. Each tree is constructed by randomly selecting data from the training set using the bagging technique. The selection of features is meticulously done using methods like Information Gain Ratio criterion or Gini Index [32].

The random forest tree is grown by first selecting a sample from the training data. Then, a decision tree is built from the sample data by repeatedly selecting a number of features from the sample, selecting the best split, and split the node into two branches. These steps are repeated until an interation threshold is reached [33].

c. Naïve Bayes Classifier (NBC)

Naive Bayes classifier uses the basic principle of Bayes theorem [34]. For a datapoint x, the probability of the data is assigned class k (C_k) is expressed as:

$$p(C_k|x) = \frac{p(C_k)p(x|C_k)}{p(x)} \quad (2)$$

In case of continuous data, the likelihood is assumed to have normal distribution. Thus, the likelihood is calculated using Formula 3

$$p(x|C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(v-\mu_k)^2}{2\sigma_k^2}}. \quad (3)$$

The class of a datapoint is determined by the maximum posterior, as in Formula 4.

$$\hat{y} = \underset{k\in\{1,\dots,K\}}{\operatorname{argmax}}\, p(C_k)\prod_{i=1}^n p(x_i|C_k) \quad (4)$$

d. K-Nearest Neighbor (k-NN)

k-NN differs from other classifiers in that it does not construct a model during its operation due to the lack of a formal training process. Instead, kNN operates on the principle that the class of a given data point is likely to be similar to that of neighboring data points, relying on the concept of a "neighborhood" [35]. This concept of neighborhood is defined by a distance measure, which determines the similarity between two data points, denoted as x and y, through various distance metrics like Euclidean, Manhattan, and Chebyshev, as specified in Formula 5-7.

$$d(x,y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (5)$$

$$d(x,y) = \sum_{i=1}^n |x_i - y_i| \quad (6)$$

$$d(x,y) = \max_i(|x_i - y_i|) \quad (7)$$

kNN determines a category of a test data by discovering k nearest neighbor after calculating the distance of the test data to each data in the training data. The class is determined by the majority class of the k nearest neighbors.

e. Multi Layer Perceptron (MLP)

MLP is a neural network configuration comprising a minimum of three layers. These layers consist of one designated input layer for processing input data, one or more hidden layers, and an output layer responsible for executing calculations and non-linearization. The introduction of non-linear computation is achieved through the utilization of an activation function [36]. A depiction of an MLP featuring one input layer accommodating 4 features, a hidden layer comprising 2 neurons, and an output layer containing 3 output neurons can be observed in Figure 2.

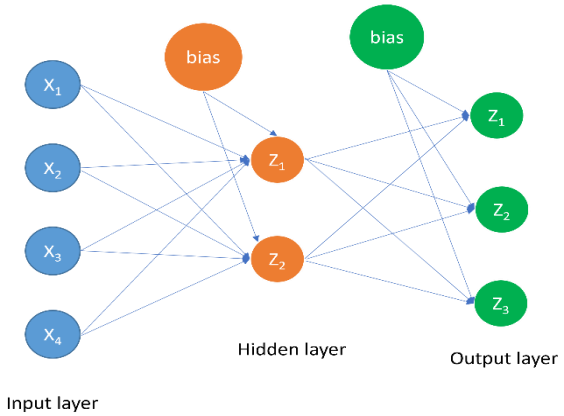


Figure 2. Illustration of an MLP network

2.4 K-Fold Cross Validation

After developing a model, it is crucial to evaluate its generalizability by assessing its performance on unseen test data. Cross-validation is a commonly used technique for this purpose. This approach involves dividing the dataset into two subsets, with one utilized for training the model and the other for testing its performance [37].

K-fold cross-validation divides the dataset into *k* equal partitions. One portion is allocated for testing while the others are for training. This process is repeated *k* times and the performance of each iteration is averaged.

2.5 Evaluation Metrics

Various methods are available for evaluating the performance of a classifier. The primary metrics used for this purpose include precision, recall, and F1 score. These metrics are derived from comparing the classifier's output with the true class labels. This comparison can be represented in a tabular form known as a confusion matrix, as depicted in Figure 3.

	Actual Positive	**Actual Negative**
Predicted Positive	True Positive (TP)	False Positive (FP)
Predicted Negative	False Negative (FN)	True Negative (TN)

Figure 3. Illustration of Confusion Matrix

Precision is determined through a confusion matrix, denoting the proportion of accurately predicted positive instances out of all predicted positive instances. In contrast, recall represents the proportion of actual positive instances correctly identified as positive. The F1 score, on the other hand, is calculated as the harmonic mean of precision and recall. These metrics, including precision, recall, accuracy, and F1 score, are mathematically represented by Formula 8-11 [38].

$$Precision = \frac{TP}{TP+FP} \quad (8)$$

$$Recall = \frac{TP}{TP+FN} \quad (9)$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (10)$$

$$F1 = \frac{2.Precision.Recall}{Precision+Recall} \quad (11)$$

The evaluation metrics can be extended to multiclass evaluation by introducing macro-averaging. Before peforming the averaging, the TP, FP, and FN are calculated based on One-vs-Rest approach, where virtually *n* classifiers were developed to classify each class. After that, the Precision, Recall, Accuracy, and F1 score are calculated using macro-averaging as follows [39]:

$$Precision_{macro} = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i+FP_i} \quad (12)$$

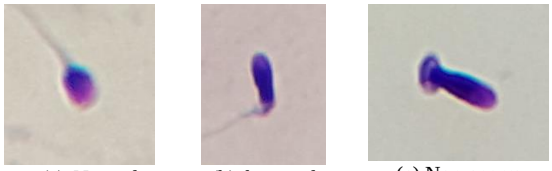
$$Recall_{macro} = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i+FN_i} \quad (13)$$

$$Accuracy_{macro} = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i+TN_I}{TP_i+TN_i+FP_i+FN_i} \quad (14)$$

$$F1_{macro} = \sum_{i=1}^{N}\frac{2.Precision_i.Recall_i}{Precision_i+Recall_i} \quad (15)$$

2.6 Dataset

This study utilizes the Sperm Morphology Image Data Set (SMIDS) [40] which consists of 3000 images of individual sperm cells categorized based on their morphology. The dataset includes three groups: normal sperm, abnormal sperm, and non-sperm. Each category's sample is illustrated in Figure 4. The distribution of images within each class is evenly balanced, as detailed in Table 2.



(a) Normal (b)abnormal (c) Non-sperm

Figure 4. Sample image for each category

Table 2. Data distribution per class.

Class	Number of data
Normal	1021
Abnormal	1005
Non-sperm	974

2.7 Experiment Setting

The proposed method is implemented using Orange [41] data mining toolbox version 3.37.0. For extracting features from image, image embedding based on SqueezeNet is used. SqueezeNet-based image embedding produces a feature vectors of 1000 elements. The network was trained using ImageNet dataset. In order to have fair evaluation, the training and testing scheme is set to 10-fold cross-validation. Hyperparameters for each machine learning model are available in Table 3 - Table 6.

Table 3. Hyperparameters for SVM.

Hyperparameter	Value
Cost	1.00
Epsilon	0.1
Kernel	RBF
Iteration limit	100

Table 4. Hyperparameters for Random Forest.

Hyperparameter	Value
Number of trees	100
Min subset split	5

Table 5. Hyperparameters for MLP.

Hyperparameter	Value
Neurons in hidden layer	128
Activation	ReLu
Solver	Adam
Max iteration	200

Table 6. Hyperparameters for kNN.

Hyperparameter	Value
Number of neighbors	5
Metric	Euclidean
Weight	Uniform

3 RESULTS AND DISCUSSION

3.1 Image Color Type

The initial study aimed to investigate the impact of the number of color channels on the performance of the classification algorithm. To achieve this objective, a SVM classifier was trained on both RGB and grayscale images, and the outcomes were assessed through a 10-fold cross-validation process.

Table 7 presents a performance analysis of SVM trained on both RGB and grayscale images. The results indicate that utilizing RGB images enhances the performance of SVM classifiers across all evaluation metrics. Nevertheless, the improvement is marginal, amounting to only 3.91%. This modest enhancement could be attributed to the adequate visibility of sperm cell defects in grayscale images. In cases where superior performance is required, it is advisable to employ RGB images. Conversely, if speed is of utmost importance, grayscale images represent a more suitable alternative.

Table 7. Classification performance on RGB dan grayscale image

Metric	Grayscale	RGB
Precision	0.749	0.773
Recall	0.741	0.771
Accuracy	0.741	0.771
F1	0.742	0.771

3.2 Comparison Between Classification Algorithms

A second experiment was conducted to determine the optimal method for categorizing defects in images of sperm cells. Five traditional machine learning algorithms, namely SVM, random forest, MLP, naive bayes, and kNN, were utilized for training and assessment through 10-fold cross-validation on RGB images. Each algorithm underwent training and evaluation using identical data splits to ensure consistency.

Table 8 presents a comparative analysis of the effectiveness of individual algorithms in the classification of the sperm data. The MLP algorithm demonstrated superior performance, as evidenced by achieving the highest F1 score. Among the algorithms assessed, only MLP and random forest were able to attain F1 scores exceeding 80%. MLP emerges as a suitable choice for this task due to its capacity to model non-linear functions and make accurate estimations when provided with a hidden layer containing an adequate number of nodes. Moreover, the dataset comprises 3000 images featuring 1000 characteristics, a combination robust enough to prevent both overfitting and underfitting. Furthermore, the simplicity of MLP's architecture serves to mitigate the risks associated with the vanishing gradient issue.

Table 8. Performance comparison on several classifiers

Algorithm	Precision	Recall	Accuracy	F1
SVM	0.773	0.771	0.771	0.771
RF	0.821	0.814	0.814	0.816
MLP	**0.854**	**0.854**	**0.854**	**0.854**
NBC	0.760	0.755	0.755	0.757
kNN	0.802	0.788	0.788	0.790

Table 9 presents the classification accuracy of each class in the dataset utilizing MLP. Based on the F1 score, the classifier displayed optimal performance in distinguishing non-sperm cells. Non-sperm cells exhibit distinct shapes that are visually distinguishable from sperm cells, allowing the classifiers to identify them accurately. Conversely, discerning between normal and abnormal cells is more challenging, as the distinctions may involve the presence or shape of specific parts of the sperm cells. Consequently, this poses increasing difficulties for classifiers in accurately recognizing these differences.

Table 9. Classification performance on each class

Class	Precision	Recall	F1
Normal	0.831	0.856	0.843
Abnormal	0.814	0.811	0.813
Non-sperm	0.921	0.895	0.908

The time required for training and testing in each classification algorithm is outlined in Table 10. In this particular scenario, the total training time only encompassed the time spent on training the classifier,

as the feature extractor utilizing image embedding was not retrained. Among the classification algorithms tested, the random forest algorithm proved to be the most time-consuming to train, requiring nearly 3 minutes to complete the training process. Conversely, the MLP algorithm, which exhibited the highest level of accuracy, took approximately one and a quarter minutes to finish training. The remaining algorithms required less than thirty seconds to complete training. It is worth noting that the training time for the kNN algorithm is zero, as kNN does not involve an actual training step. Regarding testing time, all algorithms required only a few seconds to complete the testing phase.

Table 10. Training and testing time (in second) for each classifier

Algorithm	Training Time	Testing Time
SVM	29.490	4.002
Random Forest	169.835	1.169
MLP	75.511	2.333
Naïve Bayes	5.426	1.176
kNN	0	2.332

3.3 Comparison with other methods

We also compared the result of our proposed method with those of similar approach on the same dataset. Two previous works from Ilhan et al. [42] and Yüzkat et al. [14] were taken as comparison. Ilhan et al. used SURF and MSER feature descriptors and SVM as classifiers, while Yüzkat et al. developed ensembled model from 6 CNN models.

Table 11 presents a comparison of the effectiveness of the suggested methodologies with others of a similar nature. Our methodology achieved results that were on par with those of Ilhan et al. Nonetheless, we employed a generic feature extractor as opposed to handcrafted features that may not display strong generalizability. In contrast, the method proposed by Yüzkat et al. demonstrated significant enhancements; however, this method necessitated the use of a complex combination of 6 CNN models that are costly to train. Of all the models, the least complecated one was trained for a duration of 11 hours, while some of them required twice the amount of time for training. In comparison, the combination of feature extraction and a classification algorithm in our proposed method took no more than 3 minutes at maximum to finish the training.

Table 11. Performance comparison with other methods

Method	Accuracy (%)
Ilhan et al. (SURF)	85.1
Ilhan et al. (MSER)	85.7
Yüzkat et al. (No augmentation)	66.45
Yüzkat et al. (8x augmentation)	90.2
Proposed	85.4

4 CONCLUSION

In this study, we have presented a straightforward yet efficient approach for the identification of sperm morphology in images of sperm cells. The methodology we have introduced strikes a balance between effectiveness and ease of implementation. We have attained a high accuracy rate of 85.4% with minimal training time. This finding is particularly useful for tasks requiring timely recognition of sperm morphology and for situations where hardware resources are limited, as it does not necessitate the use of a GPU and the computational process for inference is simple.

More studies are needed to determine the significant characteristics produced by image embedding, as it involves many variables. Additionally, incorporating an explainability component into the suggested approach is beneficial, particularly in the medical sector where justification for classification is often required.

5 REFERENCE

[1] J. Auger, P. Jouannet, and F. Eustache, "Another look at human sperm morphology," *Hum. Reprod.*, vol. 31, no. 1, pp. 10–23, Jan. 2016, doi: 10.1093/humrep/dev251.

[2] R. Menkveld, C. A. Holleboom, and J. P. Rhemrev, "Measurement and significance of sperm morphology," *Asian J. Androl.*, vol. 13, no. 1, pp. 59–68, Jan. 2011, doi: 10.1038/aja.2010.67.

[3] T. Kruger and K. Coetzee, "The role of sperm morphology in assisted reproduction," *Hum. Reprod. Update*, vol. 5, no. 2, pp. 172–178, Mar. 1999, doi: 10.1093/humupd/5.2.172.

[4] G. Cito *et al.*, "Sperm morphology: What implications on the assisted reproductive outcomes?," *Andrology*, vol. 8, no. 6, pp. 1867–1874, Nov. 2020, doi: 10.1111/andr.12883.

[5] S. Oehninger and T. F. Kruger, "Sperm morphology and its disorders in the context of infertility," *FS Rev.*, vol. 2, no. 1, pp. 75–92, Jan. 2021, doi: 10.1016/j.xfnr.2020.09.002.

[6] World Health Organization, *WHO laboratory manual for the examination and processing of human semen*. Geneva: World Health Organization, 2021.

[7] P. Matson, M. Kitson, and E. Zuvela, "Human sperm morphology assessment since 2010: experience of an Australian external quality assurance programme," *Reprod. Biomed. Online*, vol. 44, no. 2, pp. 340–348, Feb. 2022, doi: 10.1016/j.rbmo.2021.11.005.

[8] E. Filimberti *et al.*, "High variability in results of semen analysis in andrology laboratories in Tuscany (Italy): the experience of an external quality control (EQC) programme," *Andrology*, vol. 1, no. 3, pp. 401–407, 2013, doi: 10.1111/j.2047-2927.2012.00042.x.

[9] U. Punjabi, C. Wyns, A. Mahmoud, K. Vernelen, B. China, and G. Verheyen, "Fifteen years of Belgian experience with external quality assessment of semen analysis," *Andrology*, vol. 4,

no. 6, pp. 1084–1093, 2016, doi: 10.1111/andr.12230.

[10] C. Álvarez *et al.*, "External quality control program for semen analysis: Spanish experience," *J. Assist. Reprod. Genet.*, vol. 22, no. 11, pp. 379–387, Dec. 2005, doi: 10.1007/s10815-005-7461-2.

[11] R. Finelli, K. Leisegang, S. Tumallapalli, R. Henkel, and A. Agarwal, "The validity and reliability of computer-aided semen analyzers in performing semen analysis: a systematic review," *Transl. Androl. Urol.*, vol. 10, no. 7, Art. no. 7, Jul. 2021, doi: 10.21037/tau-21-276.

[12] R. P. Amann and D. F. Katz, "Andrology Lab Corner: Reflections on CASA After 25 Years," *J. Androl.*, vol. 25, no. 3, pp. 317–325, 2004, doi: 10.1002/j.1939-4640.2004.tb02793.x.

[13] R. P. Amann and D. Waberski, "Computer-assisted sperm analysis (CASA): Capabilities and potential developments," *Theriogenology*, vol. 81, no. 1, pp. 5-17.e3, Jan. 2014, doi: 10.1016/j.theriogenology.2013.09.004.

[14] M. Yüzkat, H. O. Ilhan, and N. Aydin, "Multi-model CNN fusion for sperm morphology analysis," *Comput. Biol. Med.*, vol. 137, p. 104790, Oct. 2021, doi: 10.1016/j.compbiomed.2021.104790.

[15] L. Spencer, J. Fernando, F. Akbaridoust, K. Ackermann, and R. Nosrati, "Ensembled Deep Learning for the Classification of Human Sperm Head Morphology," *Adv. Intell. Syst.*, vol. 4, no. 10, p. 2200111, 2022, doi: 10.1002/aisy.202200111.

[16] H. Yang *et al.*, "Multidimensional morphological analysis of live sperm based on multiple-target tracking," *Comput. Struct. Biotechnol. J.*, vol. 24, pp. 176–184, Dec. 2024, doi: 10.1016/j.csbj.2024.02.025.

[17] M. E. Kandel *et al.*, "Reproductive outcomes predicted by phase imaging with computational specificity of spermatozoon ultrastructure," *Proc. Natl. Acad. Sci.*, vol. 117, no. 31, pp. 18302–18309, Aug. 2020, doi: 10.1073/pnas.2001754117.

[18] S. Javadi and S. A. Mirroshandel, "A novel deep learning method for automatic assessment of human sperm images," *Comput. Biol. Med.*, vol. 109, pp. 182–194, Jun. 2019, doi: 10.1016/j.compbiomed.2019.04.030.

[19] J. Riordon, C. McCallum, and D. Sinton, "Deep learning for the classification of human sperm," *Comput. Biol. Med.*, vol. 111, p. 103342, Aug. 2019, doi: 10.1016/j.compbiomed.2019.103342.

[20] R. G. Tiwari, A. Misra, and N. Ujjwal, "Image Embedding and Classification using Pre-Trained Deep Learning Architectures," in *2022 8th International Conference on Signal Processing and Communication (ICSC)*, Dec. 2022, pp. 125–130. doi: 10.1109/ICSC56524.2022.10009560.

[21] D. Kiela and L. Bottou, "Learning Image Embeddings using Convolutional Neural Networks for Improved Multi-Modal Semantics," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar: Association for Computational Linguistics, 2014, pp. 36–45. doi: 10.3115/v1/D14-1005.

[22] J. Kim and Y. Kang, "Automatic Classification of Photos by Tourist Attractions Using Deep Learning Model and Image Feature Vector Clustering," *ISPRS Int. J. Geo-Inf.*, vol. 11, no. 4, Art. no. 4, Apr. 2022, doi: 10.3390/ijgi11040245.

[23] Y. Gu, Y. Xu, X. Huang, J. Yang, W. Xue, and G.-Z. Yang, "Toward Robust Histology-Prior Embedding for Endomicroscopy Image Classification," *IEEE Trans. Med. Imaging*, vol. 41, no. 11, pp. 3242–3252, Nov. 2022, doi: 10.1109/TMI.2022.3180340.

[24] Y. Xu, W. Guo, Z. Zhang, and W. Yu, "Multiple Embeddings Contrastive Pretraining for Remote Sensing Image Classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022, doi: 10.1109/LGRS.2022.3185729.

[25] Z. Ralte and I. Kar, *Learn Python Generative AI: Journey from autoencoders to transformers to large language models (English Edition)*. BPB Publications, 2024.

[26] Z. Hu, Q. Zhang, and M. He, *Advances in Artificial Systems for Logistics Engineering III*. Springer Nature, 2023.

[27] M. Berman, H. Jégou, A. Vedaldi, I. Kokkinos, and M. Douze, "MultiGrain: a unified image embedding for classes and instances," Apr. 03, 2019, *arXiv*: arXiv:1902.05509. doi: 10.48550/arXiv.1902.05509.

[28] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size," Nov. 04, 2016, *arXiv*: arXiv:1602.07360. doi: 10.48550/arXiv.1602.07360.

[29] "GitHub - alyato/CNN-models-comparison: Comparison of famous convolutional neural network models," GitHub. Accessed: Oct. 03, 2024. [Online]. Available: https://github.com/alyato/CNN-models-comparison

[30] S. Salcedo-Sanz, J. L. Rojo-Álvarez, M. Martínez-Ramón, and G. Camps-Valls, "Support vector machines in engineering: an overview," *WIREs Data Min. Knowl. Discov.*, vol. 4, no. 3, pp. 234–267, 2014, doi: 10.1002/widm.1125.

[31] M. Awad and R. Khanna, "Support Vector Machines for Classification," in *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*, M. Awad and R. Khanna, Eds., Berkeley, CA: Apress, 2015, pp. 39–66. doi: 10.1007/978-1-4302-5990-9_3.

[32] M. Pal, "Random forest classifier for remote sensing classification," *Int. J. Remote Sens.*, vol.

26, no. 1, pp. 217–222, Jan. 2005, doi: 10.1080/01431160412331269698.

[33] T. Hastie, R. Tibshirani, J. Friedman, T. Hastie, R. Tibshirani, and J. Friedman, "Random forests," *Elem. Stat. Learn. Data Min. Inference Predict.*, pp. 587–604, 2009.

[34] H. Kamel, D. Abdulah, and J. M. Al-Tuwaijari, "Cancer Classification Using Gaussian Naive Bayes Algorithm," in *2019 International Engineering Conference (IEC)*, Jun. 2019, pp. 165–170. doi: 10.1109/IEC47844.2019.8950650.

[35] S. Adinugroho and Y. A. Sari, *Implementasi Data Mining Menggunakan Weka*. Universitas Brawijaya Press, 2018.

[36] F. A. Breve, M. P. Ponti-Junior, and N. D. A. Mascarenhas, "Multilayer Perceptron Classifier Combination for Identification of Materials on Noisy Soil Science Multispectral Images," in *XX Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI 2007)*, Oct. 2007, pp. 239–244. doi: 10.1109/SIBGRAPI.2007.10.

[37] S. Yadav and S. Shukla, "Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification," in *2016 IEEE 6th International Conference on Advanced Computing (IACC)*, Feb. 2016, pp. 78–83. doi: 10.1109/IACC.2016.25.

[38] C. Sammut and G. I. Webb, *Encyclopedia of Machine Learning*. Springer Science & Business Media, 2011.

[39] M. C. Hinojosa Lee, J. Braet, and J. Springael, "Performance Metrics for Multilabel Emotion Classification: Comparing Micro, Macro, and Weighted F1-Scores," *Appl. Sci.*, vol. 14, no. 21, Art. no. 21, Jan. 2024, doi: 10.3390/app14219863.

[40] 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," *Med. Biol. Eng. Comput.*, vol. 58, no. 5, pp. 1047–1068, May 2020, doi: 10.1007/s11517-019-02101-y.

[41] J. Demšar *et al.*, "Orange: Data Mining Toolbox in Python," *J. Mach. Learn. Res.*, vol. 14, pp. 2349–2353, 2013.

[42] H. O. Ilhan, G. Serbes, and N. Aydin, "Automated sperm morphology analysis approach using a directional masking technique," *Comput. Biol. Med.*, vol. 122, p. 103845, Jul. 2020, doi: 10.1016/j.compbiomed.2020.103845.