












Automatic Unsupervised Clustering of Videos of the Intracytoplasmic Sperm Injection (ICSI) Procedure

Andrea M. Storås^{1,2} , Michael A. Riegler¹ , Trine B. Haugen³ ,
Vajira Thambawita¹ , Steven A. Hicks^{1,2} , Hugo L. Hammer^{1,2} ,
Radhika Kakulavarapu³ , Pål Halvorsen^{1,2} , and Mette H. Stensen⁴ 

¹ Department of Holistic Systems, Simula Metropolitan Center for Digital Engineering, Oslo, Norway
`andrea@simula.no`

² Department of Computer Science, Faculty of Technology, Art and Design, OsloMet - Oslo Metropolitan University, Oslo, Norway

³ Department of Life Sciences and Health, Faculty of Health Sciences, OsloMet - Oslo Metropolitan University, Oslo, Norway

⁴ Fertilitetssenteret, Oslo, Norway

Abstract. The *in vitro* fertilization procedure called intracytoplasmic sperm injection can be used to help fertilize an egg by injecting a single sperm cell directly into the cytoplasm of the egg. In order to evaluate, refine and improve the method in the fertility clinic, the procedure is usually observed at the clinic. Alternatively, a video of the procedure can be examined and labeled in a time-consuming process. To reduce the time required for the assessment, we propose an unsupervised method that automatically clusters video frames of the intracytoplasmic sperm injection procedure. Deep features are extracted from the video frames and form the basis for a clustering method. The method provides meaningful clusters representing different stages of the intracytoplasmic sperm injection procedure. The clusters can lead to more efficient examinations and possible new insights that can improve clinical practice. Further on, it may also contribute to improved clinical outcomes due to increased understanding about the technical aspects and better results of the procedure. Despite promising results, the proposed method can be further improved by increasing the amount of data and exploring other types of features.

Keywords: Unsupervised learning · Clustering · Human reproduction · Medical videos · Computer vision

1 Introduction

Infertility is defined as a disease where an individual or a couple does not succeed in becoming clinically pregnant after a period of twelve months with regular,

© The Author(s) 2022

E. Zouganeli et al. (Eds.): NAIS 2022, CCIS 1650, pp. 111–121, 2022.

https://doi.org/10.1007/978-3-031-17030-0_9

unprotected sexual intercourse [22]. Estimates suggest that about 190 million people worldwide are affected by infertility [9]. Assisted reproductive technology (ART) is used to treat infertility, and *in vitro* fertilization has been used for more than 40 years. The procedure called intracytoplasmic sperm injection (ICSI) [16] was introduced in the beginning of the 1990s, as a treatment for male factor infertility due to poor semen quality. Using this treatment, a single sperm is injected into the egg. The use of ICSI has greatly increased over the past years [1, 7].

Visual examinations of the ICSI procedure are performed to evaluate technical aspects of the procedure. Some of the critical steps during the procedure are the selection of which sperm to inject, how the immobilization of sperm is performed, the technique used for injecting the sperm into the egg and the quality of the egg. Figures 1a to 1d illustrate different stages of the procedure, as well as debris. All video frames in the figures are from the data applied in the present study. Differences in results reported after ICSI treatments are partly explained by the level of experience of the embryologist performing the procedure, but technical variations might also be important [17]. For example, videos of the ICSI procedure can be applied for training purposes and refinement of internal procedures at the fertility clinic. Detailed understanding and control of the technical procedure may lead to improved clinical outcomes as well, such as higher fertilization and pregnancy rates. However, the examination and labeling of videos are time-consuming, and it requires knowledge about the critical steps during the procedure. Furthermore, medical professionals with such knowledge are not always available for labeling medical data, which complicates the process of obtaining labeled videos of high quality. Consequently, unsupervised learning is an attractive alternative, as it allows for training artificial intelligence (AI) models without labeled data. Because the outputs from the unsupervised models are not assigned distinct labels, some type of human interpretation of the results is required, but this still requires less work than manually labeling all samples in a dataset.

In this work, we present an unsupervised clustering technique that is able to cluster video frames from the ICSI procedure into groups that represent different stages of the procedure. This can make the examination of the videos more effective, and the health personnel will save time as they can watch the critical steps directly. Further on, focusing on the relevant parts of the procedure might contribute to easier detection of possible improvements, which could lead to improved clinical outcomes. Unsupervised clustering techniques have been developed for summarization of capsule endoscopy videos [8], detection of anomalies in computed tomography (CT) scans [3], segmentation of 3D medical images [14] and to diagnose coronavirus disease (CoVID19) from medical images [13]. None of these studies apply the same clustering algorithms as in the present paper, and they do not investigate data from the field of human reproduction. Regarding the use of AI to analyze videos of the ICSI procedure, one study trained a U-Net neural network to extract video frames of the oolemma, i.e., the cell membrane of the egg, during sperm injection [10]. To our knowledge,

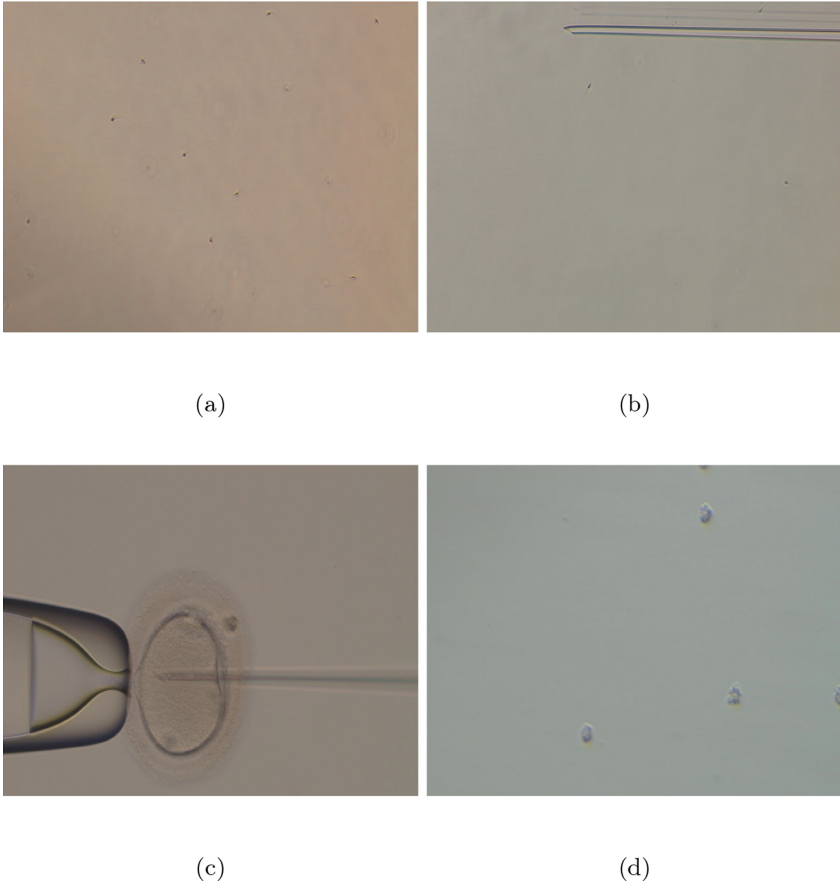


Fig. 1. Examples of video frames representing sperm selection (a), sperm immobilization (b), sperm injection (c) and debris (d). The frames arrive from the data applied in the presented work.

this is the first time unsupervised clustering has been applied to video frames of the ICSI procedure. Thus, the main contributions of this work are:

- Exploration of unsupervised learning and clustering on extracted video frames of the ICSI procedure as a tool for embryologists.
- Automatic detection of important stages in recordings of the procedure.
- Testing of the clustering method with embryologists to evaluate clinical practicability.
- The source code is provided for the implemented methodology.

In the following, Sect. 2 provides an overview of the data and methods used in this work. This is followed by a description of our experiments and a presentation of our results in Sect. 3. Next, our findings and their implications on the clinical

practice are discussed in Sect. 4. Finally, we provide a conclusion and possible future directions in Sect. 5.

2 Data and Method

Seven videos of artificial reproduction using the ICSI procedure are used in the experiments. The videos arrive from a pilot study that was conducted at Fertilitetssenteret in Oslo, Norway in 2021. Because the data is anonymized, no ethical approvals are required. The resolution is 1920×1080 , and the frame rate per second is 25 for all videos. The video length ranges from 15 seconds to more than 2 minutes. The longest video includes sperm selection, immobilization and injection, while the other videos capture one or two of the stages. All videos were captured at $200\times$ magnification with a DeltaPix camera. The ICSI procedure was performed using a Nikon ECLIPSE TE2000-S microscope connected with Eppendorf TransferMan 4m micromanipulator. The sperm cells were immobilized in $5 \mu\text{l}$ Polyvinylpyrrolidone (PVP; CooperSurgical). The clinical outcome of the procedures is not included in the analysis.

Figure 2 provides an overview of the proposed workflow for unsupervised clustering. Video frames are extracted every second from the seven videos using the OpenCV library in Python [2]. The frequency of one second is chosen in order to extract frames reflecting the video contents without losing much information. The extracted frames are passed through a convolutional neural network (CNN), ResNet50 [6], that has been pre-trained on the ImageNet data set [20]. Features are extracted from the layer preceding the output layer, resulting in 2,048 deep features per frame. Further on, dimensionality reduction with t-SNE [11] is applied on the extracted features. By reducing the dimensions of the data to two, the distribution of the video frames can easily be plotted for visual inspection, and the proposed method becomes more transparent. Moreover, dimensionality reduction has been applied prior to clustering of video frames to speed up the analysis [8]. t-SNE is chosen because it is an efficient technique for dimensionality reduction that has shown good performance on high-dimensional data points such as images [11]. When applying t-SNE, the user must specify the perplexity hyperparameter value, which can be thought of as a measure of the effective number of neighbors for each data point. Usually, the value should lie between 5 and 50 [11]. The perplexity values of 10, 15, 20 and 30 are tested for our data. The perplexity values chosen are based on the size of our dataset. The value should be smaller than the total number of samples to avoid one large cluster. On the other hand, values that are too small will result in local variations. The dimensionality reduction is evaluated by visually inspecting plots of the results, and identifying the plot with the most distinct clusters. The output from t-SNE is clustered using unsupervised clustering. Because the optimal number of clusters is not known, X-means clustering [19] is applied to determine the appropriate number of clusters. G-means clustering [5] is also tested. Both algorithms identify the optimal number of clusters in the provided data. They are wrappers around the k-means algorithm [12], and the final clusters depend on

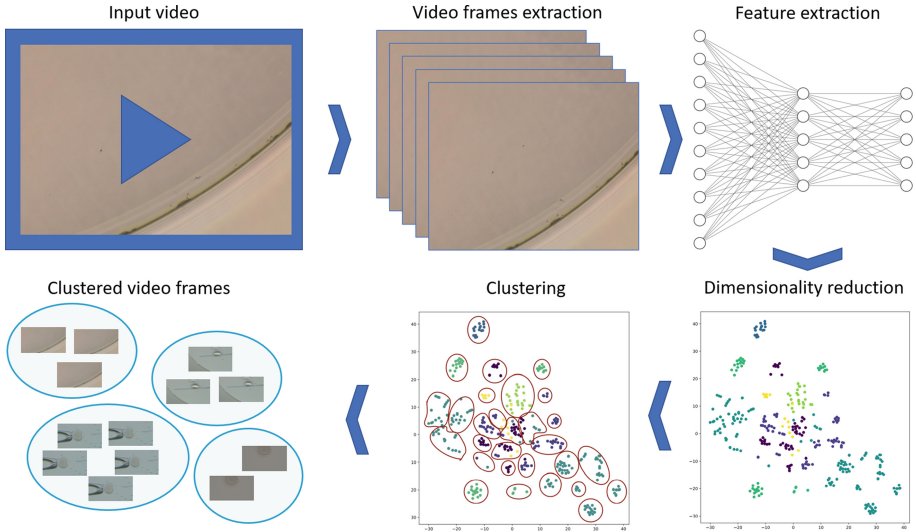


Fig. 2. The workflow for the proposed clustering method. First, video frames are extracted from videos of the ICSI procedure. The frames are then passed through a pretrained ResNet50 for extraction of deep features. The dimensionality is reduced using t-SNE before the frames are clustered using either X-means or G-means.

the cluster initialization. Consequently, the results can vary between runs even though the dataset is the same. While X-means applies the Bayesian Information Criterion to find the appropriate cluster number, G-means, on the other hand, uses a Gaussian fit. The G-means algorithm has shown higher performance than X-means when the clusters are non-spherical [5]. All code is written in Python. Pyclustering is applied for unsupervised clustering [15], and Pytorch [18] is used for extracting the deep features from the pretrained ResNet50 model [6]. The source code is publicly available online¹.

The quality of the clustering is evaluated by experienced embryologists working at Fertilitetssenteret in Oslo, Norway. The clusters are also categorized into which stage of the ICSI procedure they represent to evaluate the accuracy of the methods, but this is regarded as less important than the feedback from the embryologists.

3 Results

In total, 359 images are extracted from the seven videos. The extracted deep features are reduced to two dimensions using t-SNE. Following visual inspection, the best perplexity hyperparameter value for t-SNE is 20, leading to the most distinct clusters. The results are shown in Fig. 3. Regarding the unsupervised clustering, the X-means algorithm suggested two clusters for the data when no

¹ https://github.com/AndreaStoraas/UnsupervisedClustering_ICSI.

restrictions were set. However, this is not regarded as a sufficient number of clusters due to the variation between the frames. Consequently, the algorithms are restricted to estimating the number of clusters to lie between eight and 200. These limits are chosen to get clusters representing the variation in the dataset while not creating clusters that are too small with respect to the dataset size. When the X-means algorithm is forced to generate between eight and 200 clusters, the suggested number of clusters varies a lot for the same data set, ranging between 8 and 15 clusters. This makes it challenging to determine the appropriate number of clusters to use with the X-means algorithm. On the other hand, the G-means clustering algorithm is more stable, suggesting 29 or 30 clusters. Consequently, the clusters from the G-means algorithm are further investigated and evaluated by domain experts. The 29 clusters suggested by the G-means algorithm are indicated in Fig. 4.

The video frames in all of the 29 clusters are shown to four experienced embryologists working at Fertilitetssenteret in Oslo for evaluation of the quality of the clusters and detection of potential weaknesses of the method. An overall finding is that the clusters are dependent on the colors and the presence of edges in the frames. Moreover, two of the experts, one being a senior embryologist and the other one being a clinical embryologist, manually categorize the clusters after examination of typical examples of video frames from different clusters. Based on their feedback, the clusters are categorized into three subgroups that represent different critical stages of the ICSI procedure: sperm selection, sperm immobilization, and sperm injection. Video frames from these three subgroups can be studied more closely to inspect which sperm was selected, how it was immobilized and the technique applied when injecting the sperm into the egg. A fourth subgroup is also created for video frames containing bubbles and debris, here defined as noise.

The feedback from the embryologists is the main evaluation of the method. However, the accuracy of the clustering was also investigated as a secondary measure of performance. Based on visual inspection, 82% of the frames are automatically assigned to a cluster belonging to the same category. The categories were provided by the domain experts, as described above. The sperm selection seems to be the easiest part of the ICSI procedure to recognize. Still, some frames representing sperm immobilization were clustered together with sperm selection frames. Figures 5a and 5b show examples of video frames that were clustered as sperm selection according to the cluster categories from the domain experts. Figure 5a agrees with the cluster category, while Fig. 5b disagrees. Sperm immobilization is most difficult to recognize by the method, as all the clusters that include frames from this part also contain frames presenting sperm injection or sperm selection. Video frames that were clustered as sperm immobilization are provided in Figs. 5c and 5d. Figure 5c agrees with the cluster category, while Fig. 5d disagrees.

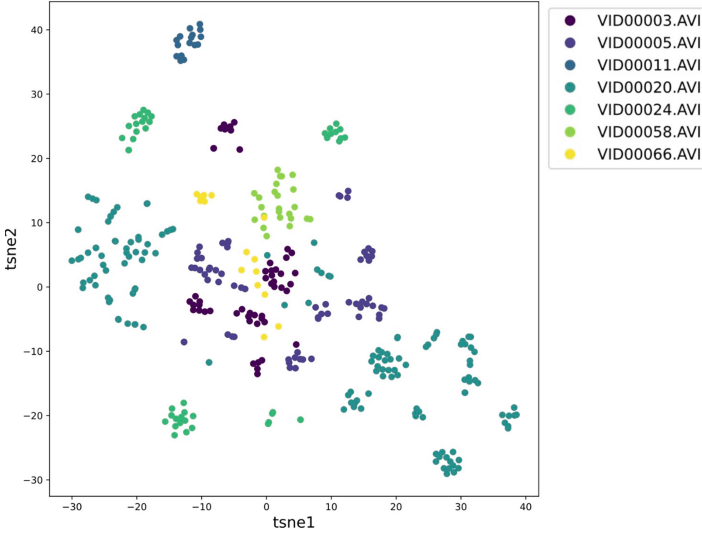


Fig. 3. Plot of the 359 images after feature extraction with a pretrained convolutional neural network and dimensionality reduction with t-SNE. The frames are colored after which video they belong to.

4 Discussion

In this work, we show that unsupervised clustering can be applied for extracting video frames from different stages of the ICSI procedure. Despite promising results, there are some limitations to be discussed. First, the proposed technique is negatively affected by the colors and edges present in the frames. Indeed, colors and edges can vary between different dishes and droplets. To make the method more robust, features that do not rely on these properties will be explored for future experiments. To reduce the variation in colors, the frames can be converted into grayscale before they are analyzed. Further on, global features such as Tamura features [21] or fuzzy color and texture histogram [4] can be applied for less dependency on the presence of edges.

Moreover, our data set included seven videos from the same fertility center. Consequently, it is not known how well they generalize to larger data sets or other clinics. Since the method is sensitive to variations in colors and edges, the performance could be affected by the resolution, light and type of camera applied during the recording of the procedure. A follow-up study is planned with more videos, as well as information about the outcome, such as fertilization status, egg degeneration rate, embryo quality, embryo development, implantation and pregnancy rates.

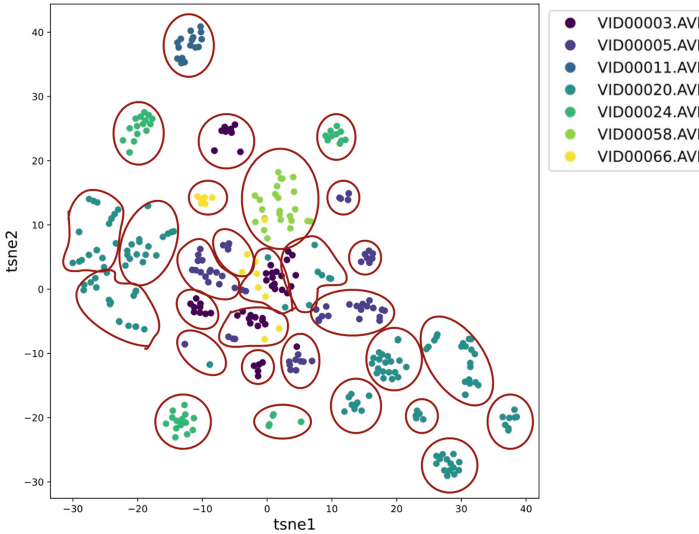


Fig. 4. Results from unsupervised clustering. The 29 clusters identified by the G-means algorithm for the 359 video frames are indicated with circles.

Our results suggest that the sperm selection stage was easiest to detect with the proposed method. The sperm selection stage does not contain any needles or eggs, which might explain why this stage is more easily separated from the other stages. The stage that was most difficult to separate was the immobilization of sperm. This could be because the sperm cells are relatively small compared to the size of the injection needle, as well as the presence of noise in the frames, making it challenging to distinguish features from these frames from features representing only noise or sperm injection.

After manual inspection of the clusters, 82% of the video frames were placed in a cluster representing the same category, as defined by domain experts. Some frames were placed in clusters representing a different category, meaning that the medical experts will encounter some frames that are not appropriate for a given stage of the ICSI procedure. Nevertheless, since most of the frames in each cluster are similar, the clusters would still be useful for a more efficient examination of the ICSI procedure. With the additional experiments suggested above, the percentage of video frames disagreeing with their cluster category might also be further reduced. Further on, the labeled clusters from our experiments can potentially be used as labels in a supervised or semi-supervised learning framework in order to categorize new video frames.

Normally, the ICSI procedure is evaluated through live observation at the clinic. Alternatively, recordings of the procedure can be watched and labeled manually. According to the senior embryologist at Fertilitetscenteret, our method proposes a more time-efficient way to improve training and quality assessment of the ICSI procedure. Because this potentially leads to improved results of the

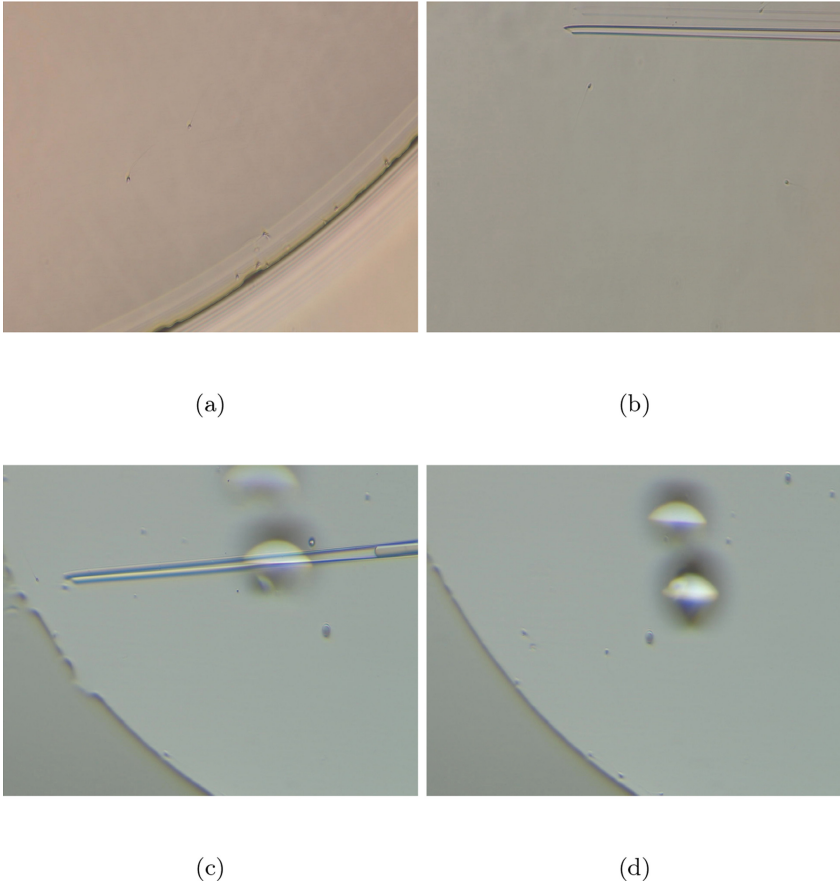


Fig. 5. Examples of video frames in clusters representing sperm selection (a, b) and sperm immobilization (c, d), according to the cluster categories provided by domain experts. Frames **a** and **c** agree with their assigned cluster labels, while **b** and **d** are video frames that were placed in clusters with a different category.

procedure, the clinical outcome, such as higher fertilization and pregnancy rates, might also improve. Finally, it could benefit couples suffering from infertility as well as the healthcare personnel performing the treatment.

5 Conclusion

In this paper, we present an unsupervised method for clustering of video frames of the popular *in vitro* fertilization technique called ICSI. Deep features are extracted from the video frames before dimensionality reduction is applied. Clustering is then performed on the resulting data points. The clusters are evaluated by experienced domain experts, and the findings are discussed. The source code for the proposed method is available online.

In conclusion, our method is able to separate video frames into different stages of the ICSI procedure. This could be valuable in the fertility clinic in order to analyze ICSI videos more efficiently for training purposes, internal quality control and refinement of internal procedures. Further on, it might improve the results after treatments with ICSI, which in turn could lead to improved clinical outcomes such as higher fertilization and pregnancy rates.

For future work, we plan to experiment with features that are less affected by the change of color and the presence of edges in the video frames. We will also use a larger data set containing an increased number of videos preferably from different clinics to see if the method can be further improved.

References

1. Boulet, S.L., Mehta, A., Kissin, D.M., Warner, L., Kawwass, J.F., Jamieson, D.J.: Trends in use of and reproductive outcomes associated with intracytoplasmic sperm injection. *JAMA* **313**(3), 255–263 (2015). <https://doi.org/10.1001/jama.2014.17985>
2. Bradski, G.: The OpenCV library. *Dr. Dobb's J. Softw. Tools* **120**, 122–125 (2000)
3. Chaira, T.: A novel intuitionistic fuzzy C means clustering algorithm and its application to medical images. *Appl. Soft Comput.* **11**(2), 1711–1717 (2011). <https://doi.org/10.1016/j.asoc.2010.05.005>
4. Chatzichristofis, S.A., Boutalis, Y.S.: FCTH: fuzzy color and texture histogram - a low level feature for accurate image retrieval. In: 2008 Ninth International Workshop on Image Analysis for Multimedia Interactive Services, pp. 191–196 (2008). <https://doi.org/10.1109/WIAMIS.2008.24>
5. Hamerly, G., Elkan, C.: Learning the k in k-means. *Adv. Neural. Inf. Process. Syst.* **16**, 281–288 (2004)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016)
7. Wyns, C., et al.: ART in Europe, 2017: results generated from European registries by ESHRE. *Hum. Reprod. Open* **2021**(3) (2021). <https://doi.org/10.1093/hropen/hoab026>
8. Iakovidis, D.K., Tsevas, S., Maroulis, D., Polydorou, A.: Unsupervised summarisation of capsule endoscopy video. In: 2008 4th International IEEE Conference Intelligent Systems, vol. 1, pp. 3–15–3–20 (2008). <https://doi.org/10.1109/IS.2008.4670414>
9. Inhorn, M.C., Patrizio, P.: Infertility around the globe: new thinking on gender, reproductive technologies and global movements in the 21st century. *Hum. Reprod. Update* **21**(4), 411–426 (2015). <https://doi.org/10.1093/humupd/dmv016>
10. Jain, R., et al.: P-280 changes in oolemma height during ICSI injection on day 0 is associated with day 5–6 blastocyst formation. *Hum. Reprod.* **36**(Supplement_1), i263 (2021)
11. Van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**(11), 2579–2605 (2008)
12. MacQueen, J., et al.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 281–297. University of California (1967)

13. Mittal, H., Pandey, A.C., Pal, R., Tripathi, A.: A new clustering method for the diagnosis of CoVID19 using medical images. *Appl. Intell.* **51**(5), 2988–3011 (2021). <https://doi.org/10.1007/s10489-020-02122-3>
14. Moriya, T., et al.: Unsupervised segmentation of 3D medical images based on clustering and deep representation learning. In: Gimi, B., Krol, A. (eds.) *Medical Imaging 2018: Biomedical Applications in Molecular, Structural, and Functional Imaging*, vol. 10578, pp. 483–489. International Society for Optics and Photonics, SPIE (2018). <https://doi.org/10.1117/12.2293414>
15. Novikov, A.: PyClustering: data mining library. *J. Open Source Softw.* **4**(36), 1230 (2019). <https://doi.org/10.21105/joss.01230>
16. Palermo, G., Joris, H., Devroey, P., Van Steirteghem, A.: Pregnancies after intracytoplasmic injection of single spermatozoon into an oocyte. *Lancet* **340**(8810), 17–18 (1992). [https://doi.org/10.1016/0140-6736\(92\)92425-F](https://doi.org/10.1016/0140-6736(92)92425-F). Originally published as Volume 2, Issue 8810
17. Palermo, G.D., Neri, Q.V., Rosenwaks, Z.: To ICSI or not to ICSI. In: *Seminars in Reproductive Medicine*, vol. 33, pp. 92–102. Thieme Medical Publishers (2015). <https://doi.org/10.1055/s-0035-1546825>
18. Paszke, A., et al.: Pytorch: an imperative style, high-performance deep learning library. In: Wallach, H., Larochelle, H., Beygelzimer, A., d’Alché-Buc, F., Fox, E., Garnett, R. (eds.) *Advances in Neural Information Processing Systems*, vol. 32, pp. 8024–8035. Curran Associates, Inc. (2019)
19. Pelleg, D., Moore, A.W.: X-means: extending k-means with efficient estimation of the number of clusters. In: *ICML*, vol. 1, pp. 727–734 (2000)
20. Russakovsky, O., et al.: ImageNet large scale visual recognition challenge. *Int. J. Comput. Vision* **115**(3), 211–252 (2015). <https://doi.org/10.1007/s11263-015-0816-y>
21. Tamura, H., Mori, S., Yamawaki, T.: Textural features corresponding to visual perception. *IEEE Trans. Syst. Man Cybern.* **8**(6), 460–473 (1978). <https://doi.org/10.1109/TSMC.1978.4309999>
22. WHO: International classification of diseases, 11th revision (ICD-11) (2018). <https://icd.who.int/en>

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

