Problems of identification in information systems

UDC 519.618 **doi: 10.20998/2522-9052.2018.3.01**

O. Akhiiezer, O. Dunaievska, I. Serdiuk, S. Spivak

National Technical University «Kharkiv Polytechnic Institute», Kharkiv, Ukraine

MACHINE LEARNING METHODS APPLICATION FOR SOLVING THE PROBLEM OF BIOLOGICAL DATA ANALYSIS

According to statistics, every fifth married couple is faced with the inability to conceive a child. Male germ cells are very vulnerable, and the growing number of cases of male infertility confirms that in today's world there are many factors that affect the activity of spermatozoa and their number. But the important thing is not so much their quantity, but quality. The spermogram is an objective method of laboratory diagnosis, which allows to accurately assess the man's ability to fertilize by analyzing ejaculate for a number of key parameters. Only a spermogram can answer the question of a possible male infertility and the presence of urological diseases. When constructing spermograms, it is important to determine not only the number of good spermatozoa, but also their morphology and mobility. Therefore, research and improvement of some stages of spermogramm is the purpose of the study. This article addresses the problem of classification of spermatozoa in good and bad ones, taking into account their mobility and morphology, using methods of machine learning. In order to implement the first stage of machine learning (with a teacher) in the graphic editor, educational specimens (training sample) were created. The training was implemented by three methods: the method of support vector machine, the logistic regression and the method of K - the nearest neighbors. As a result of testing, the method K - the nearest neighbors is chosen. At the testing stage, a sample of 15 different spermatozoa was used in different variations of rotation around their axis. The test sample did not contain specimens from the training sample and was formed taking into account the morphological characteristics of the spermatozoa, but did not copy them from the training sample. At the final stage of study, the program's functioning was tested on real data.

Keywords: machine learning; spermogram; morphology; mobility; pattern recognition; binary classification; method K– the nearest neighbors.

Introduction

Male and female infertility is a problem that is relevant all over the world. According to statistics, every fifth married couple is faced with the inability to conceive a child. Male germ cells are very vulnerable, and the growing number of cases of male infertility confirms that in the modern world there are many factors that affect both the activity of spermatozoa and their number [1, 4]. But when fertilizing the ovule, not only the quantity and mobility of the spermatozoa, but also their morphology, that is, the appearance, is of paramount importance. Only the spermatozoa with normal shape move in a straightforward manner with the required speed and can give birth to a new life. Different anomalies of the body and tail of spermatozoa reduce the chances of conceiving naturally.

Statement of the problem

It is precisely to assess the quality of biological data that the spermogram exists. The applied form of spermogram is based on the standards of the World Health Organization. The document contains the following graphs listing the main parameters that are evaluated during the analysis: the volume of the ejaculate; acid-base balance; definition of spermatozoa antibodies; leukocytes; concentration of spermatozoa; total spermatozoa count; mobility; morphology; share of live spermatozoa.

Some other physical and chemical properties are also studied: appearance, hue, viscosity, speed of self-expansion. Sharp deviations from normal values indicate a pathology that reduces or completely eliminates reproductive function. And if most of the parameters can be estimated without the use of an ECM, only with chemical reagents, others, such as concentration and quantity - are evaluated using a special device - hemocytometer, then the latter can be improved with the help of special equipment and software.

Currently, there are only a few systems for the recognition and testing of spermatozoa, but all of them are based on the CASA (Computer Assisted Sperm Analysis) [8, 9]. Let's consider the two most common: IVF 1 and SCA.

The essence of the IVF 1 system is as follows. When computer analysis of sperm, the computer identifies and tracks every sample that is observed in the microscopic field. For a fraction of a second the path through which the sperm passed can be traced, and many different parameters can then be calculated with a high degree of accuracy. Fig. 1 shows the functionning of the IVF system 1: shaky lines are the path of individual spermatozoa. The dots mark the non-moving spermatozoa.

But this computer system for determining morphology is only at the testing stage and can conduct research only on the basis of Kruger's criterion [2, 3].

SCA The SCA® CASA system for sperm analysis (Fig. 2) allows for accurate, repeat and automatic evaluation of the following sperm parameters: mobility, concentration, morphology, DNA fragmentation, viability and acrosomal response.

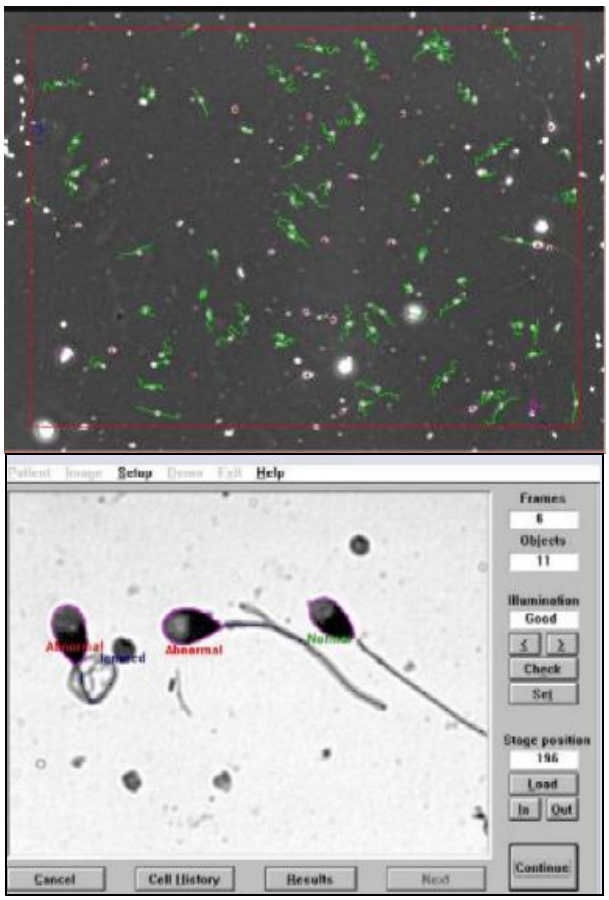

Fig. 1. The functionning of the IVF system 1

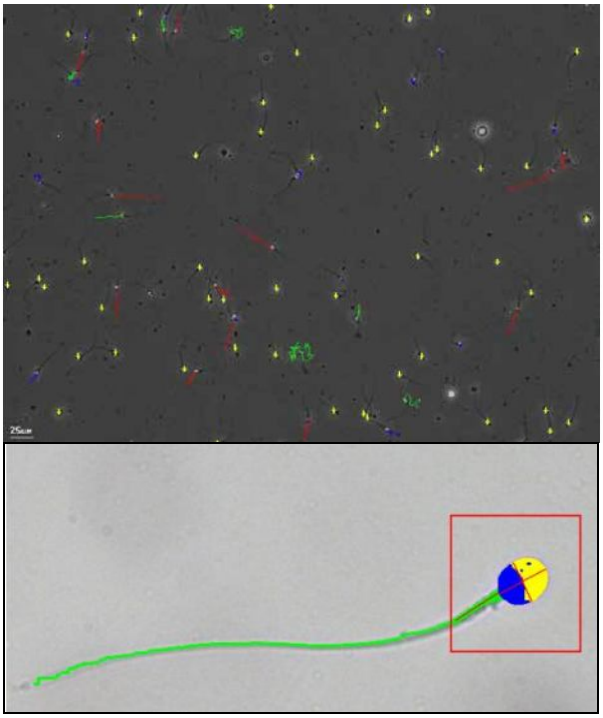

Fig. 2. The functionning of the SCA system for the evaluation of morphology

Spermatozoa are registered as normal or abnormal, according to established criteria: W.H.O., Strict Kruger-Tygerberg, David modified, adaptive or individual. At the same time, the system evaluates the size of the acrosomes, the head, the middle, the tail and the presence of vacuoles. However, the analysis of biological materials using this system is very costly and not always fast.

The classification of spermatozoa in the normal and pathological forms by methods of machine learning, the allocation of their morphological features, the determination of the trajectory of motion and the calculation of distribution density will undoubtedly help scientists and biologists to improve and facilitate the research methods and improve the accuracy of the calculation for the compilation of spermogram. Therefore, there is a need to develop a system for analyzing biological data, which, on the one hand, would work in real time, and on the other hand, its operation would not lead to large financial costs.

Solving the problem

Taking into account the above-mentioned, it was decided to solve the problem of mobility by means of computer vision, and the problem of the establishment of morphological deviations - by means of machine learning [6, 7].

In order to determine how the spermatozoa look in the norm, it is necessary to correlate them with certain diagnostic criteria [5]. Normal spermatozoon consists of an oval head and a long, curved tail. Abnormal forms are characterized by very large or, conversely, small heads, double tails, irregular head shape, and others.

Initially, the input data were screenshots from a real video provided by the biological center, obtained using a microscope. Since there is a plurality of biological objects present on the video, and a single specimen is required for the training sample, each spermatozoon has been allocated with the additionally developed application (Fig. 3) and the resulting image was converted into binary format: black and white.

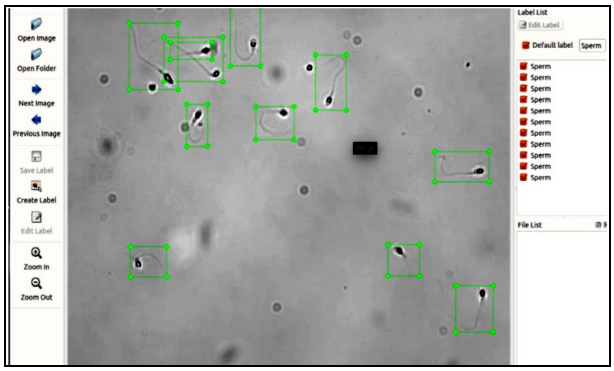

Fig. 3. Application for selecting individual objects

Since each instance has a different contrast level compared to the background, a single threshold value was selected. As a result, the data was not suitable for analysis and training, since they looked completely different from the original frames (Fig. 4).

Fig. 4. Initial data with a dimension of 72 by 72 pixels for the training sample

The figure shows that the data obtained are unsatisfactory for processing, analysis and classification. Based on this, it was decided to form perfect, close to real data in order to carry out a learning process on them and to see if further research is appropriate. Acquainted with the morphology and peculiarities of real spermatozoa, educational specimes were created in the graphic editor.

At the first stage of machine learning, a normal spermatozoon is considered as a reference, and a double one - as a defect, since it is this group that has the most pronounced deviations for analysis. The data were 1 normal and 2 bad spermatozoa (two tails and two heads), each of which was turned around its axis in a cycle with an increase of 1 degree (Fig. 5).

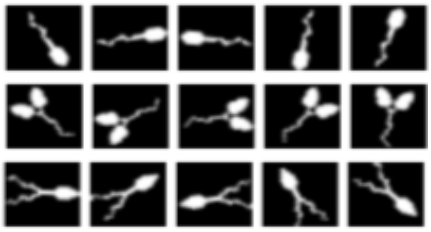

Fig. 5. Initial view of the data of the study sample (1077 pieces)

The data were received by the program, not as an image, but were presented as a code (Fig. 6).

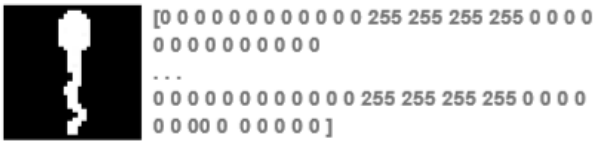

Fig. 6. Software representation of images

Since the classification problem is solved, and given the nature of the data for training, we will consider three methods of machine learning: support vector machine (SVM), logistic regression (LR) and K-the nearest neighbors (KNN).

The functionning of the selected three methods has been tested on the training sample. It turned out that the method K - the nearest neighbors works better than others and has a sufficiently high speed of training. The results of the three teaching methods are shown in Fig. 7-9.

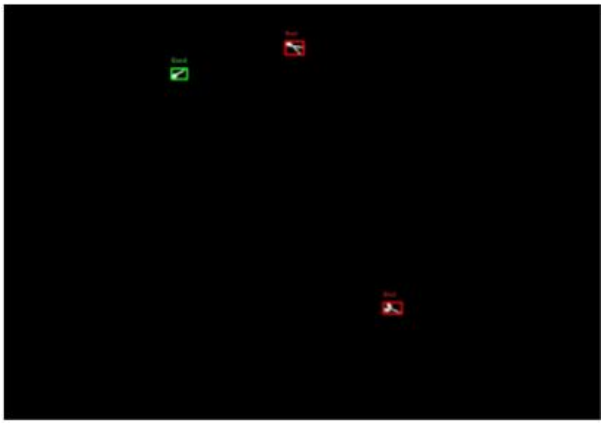

Fig. 7. The result of the program using the KNN method

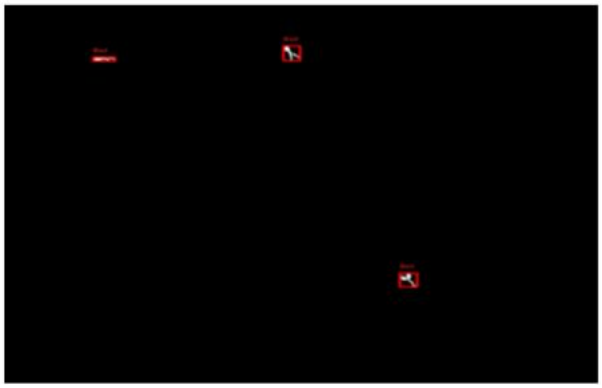

Fig. 8. The result of the program using the SVM method

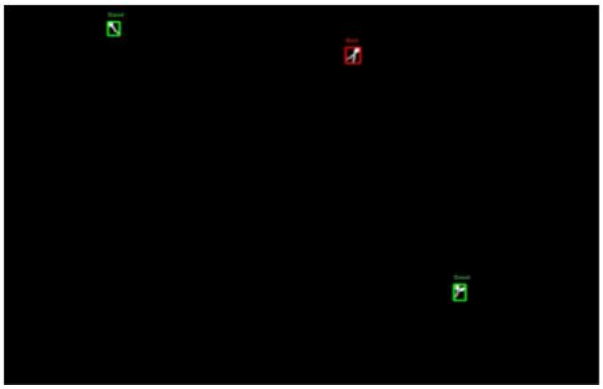

Fig. 9. The result of the program using the LR method

At the second stage of machine learning a sample for further study was formed. Each specimen of the data for the training sample was placed on a large black coordinate plane measuring 100 by 100 pixels at its various points. In total 34560 elements of the training sample were received (Fig. 10). Since the spermatozoa were stored in blocks (good-bad-bad), the first 11520 pieces were labeled as good, and the remaining (2/3) were bad.

Fig. 10. The final version of the training sample

For testing, a sample of 15 different spermatozoa was used in different variations of rotation around their axis. The test sample did not contain copies from the training one and was formed taking into account the morphological characteristics of the spermatozoa, but did not copy them from the training sample (Fig. 11).

Fig. 11. Test sample measuring 30 by 30 pixels

Next, three spermatozoa were randomly selected (1 good and 2 bad) and were placed on a black background with the dimension of 720 by 720 pixels. On this background, the samples also rotated around their axis and also directly moved around the background in different directions on a certain trajectory. The video was artificially added to the rubbish of three types of different sizes and shapes.

The final picture for the testing was as follows: one good and two bad spermatozoa with a different angle of rotation and in different points on the coordinate plane moved in a random direction with a certain speed. Artifacts simulating biodegradable garbage in the amount of 9 pieces, three of each type, also at different random points on the coordinate plane (Fig. 12).

Fig. 12. A screenshot of the video that served as a platform for testing

The result of the program functionning on artificially simulated data using the KNN algorithm is presented in Fig. 13.

In the figure we can see that the main task of recognizing biological images is realized: the spermatozoon is framed with the help of the framework of the corresponding color. White, if it meets the necessary requirements and is recognized as good or otherwise, it is placed in the frame of grey color, which corresponds to its unsuitability for further biological tasks. Above the appropriate object is the inscription "good", which corresponds to a good specimen and "bad". For each of the spermatozoa the head is defined and it stands out in dot, and if the object has two heads, then each one stands out in its circle. The trajectory of the spermatozoa movement is represented by a white line, which originates from the center of the head. The distance traveled is always written under the object, it is calculated from the starting point of motion. In the upper right corner, the counter counts all spermatozoa on the plane and records these values in two categories: good and bad. Artifacts are not affected, but simply ignored as an object that does not carry any useful information.

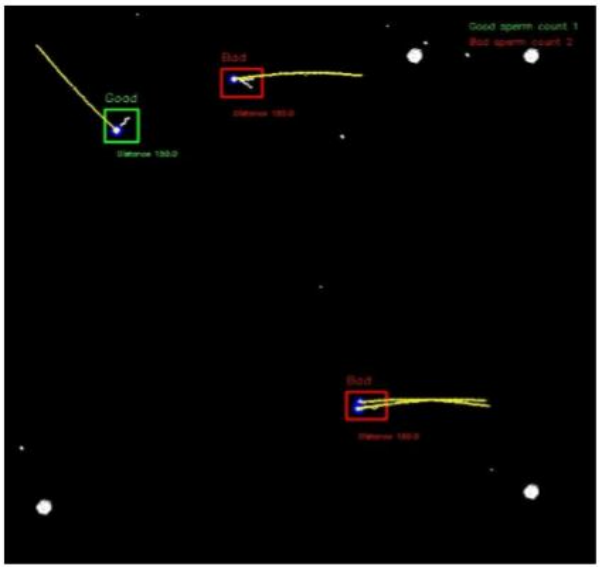

Fig. 13. The result of the program with the KNN method

At the final stage of the training the program was tested on real data. In Fig. 14 shows the work and results of the program on one of the real-time live stop-frame.

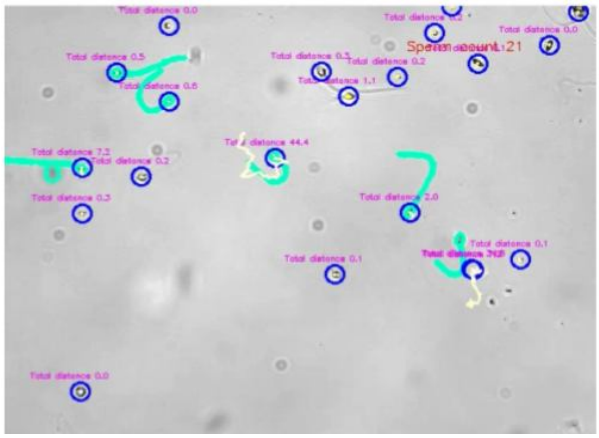

Fig. 14. Representation of the full functionality of the program on real data

Analyzing the results of the program, we can make a number of certain conclusions. The program shows good results for all the initial requirements for both artificially simulated data and real data that were provided by the biological center.

Conclusions

The task of displaying the trajectory, analysis of the traversed path for each individual sample is fully executed on both types of data, as well as the allocation of the head. Artifacts are not taken into consideration in both cases, the calculation of all biological objects is also successfully performed, which in the future will allow finding the concentration of spermatozoa for spermogram. As for the differences, on the real data we partly managed to morphologically allocate tails. On artificially modeled, we managed to achieve the classification of biological objects in normal and defective, although the study was conducted on the most notable morphological pathologies.

REFERENCES

1. Danilov, V.V. and Lelchuk, S.A. (2013), “Quantitative evaluation of spermograms in men based on interval scale”, *Andrology and Genital Surgery*, No.1, pp. 44–48.
2. Kruger, T.F., Acosta, A.A. and Simmons, K.F. (1987), “New method of evaluating sperm morphology with predictive value for human in vitro fertilization”, *Urology*, 30, pp. 248–251.
3. Kruger, T.F., Menkveld R and Stanger F.S.H. (1986), “Sperm morphologic features as a prognostic factor in in vitro fertilization”, Fertil Steril, 46, pp. 1118–1123.
4. World Health Organization, Department of Reproductive Health and Research (2010), *WHO laboratory manual for the examination and processing of human semen*, 5th edition, 234 p.
5. World Health Organization (2001), *WHO guidelines on laboratory research of human ejaculate and interaction of spermatozoa with cervical mucus*, 4th edition, Publishing house "MedPress", Moscow, 144 p.
6. Theodoridis, Sergios and Koutroumbas, Konstantinos (2009), *Pattern Recognition*, 4th Edition, Academic Press, 984 p.
7. Poreva A.S. and Karpliuk, E.S. (2017), “Methods of machine learning for the study of lung sounds”, *Biomedical instruments and systems*, No. 22 (6), pp. 41–46.
8. Talarczyk-Desole, J., Berger, A., Taszarek-Hauke, G., Hauke, J., Pawelczyk, L. and Jedrzejczak P. (2017), “Manual vs. computer-assisted sperm analysis: can CASA replace manual assessment of human semen in clinical practice?”, Ginekologia Polska, No. 88 (2), pp. 56–60.
9. (2014), *Systems M-AD. SCA - Sperm Class Analyzer v. 4.1*, polish version, Microptic - Automatic Diagnostic Systems.

Received (Надійшла) 11.06.2018
Accepted for publication (Прийнята до друку) 22.08.2018

Застосування методів машинного навчання
для вирішення задачі аналізу біологічних даних

О. Б. Ахієзер, О. І. Дунаєвська, І. В. Сердюк, С. В. Співак

За статистикою, кожна п'ята подружня пара стикається з неможливістю зачаття дитини. Чоловічі статеві клітини дуже вразливі, зростаюче число випадків чоловічого безпліддя підтверджує, що в сучасному світі дуже багато чинників, які впливають і на активність сперматозоїдів і на їх кількість. Та важливою є не стільки їх кількість, скільки якість. Спермограма є об'єктивним методом лабораторної діагностики, що дозволяє максимально точно оцінити здатність до запліднення чоловіка, проаналізувавши еякулят за рядом найважливіших параметрів. Тільки спермограма здатна відповісти на питання про можливе чоловіче безпліддя та про наявність урологічних захворювань. При побудові спермограми, важливо визначати не тільки кількість добрих сперматозоїдів, але й їх морфологію та рухливість. Тому дослідження та вдосконалення деяких етапів спермограми і є метою дослідження. У даній статті вирішується задача класифікації сперматозоїдів на добрі та погані, з урахуванням їх рухливості та морфології, із застосуванням методів машинного навчання. Для реалізації першого етапу машинного навчання (з вчителем) у графічному редакторі були створені навчальні екземпляри (тренувальна вибірка). Навчання було реалізована трьома методами: методом опорних векторів, логістична регресія та метод К – найближчих сусідів. За результатами тестування обрано метод К – найближчих сусідів. На етапі тестування використовувалася вибірка з 15 різних сперматозоїдів в різних варіаціях обертання навколо своєї осі. Тестова вибірка не містила примірників з тренувальної вибірки і була сформована з урахуванням морфологічних особливостей сперматозоїдів, але не копіювала їх з тренувальної вибірки. На завершальному етапі навчання роботу програми було протестовано на реальних даних.

Ключові слова: машинне навчання; спермограма; морфологія; рухливість; розпізнавання образів; бінарна класифікація; метод К-найближчих сусідів.

Применение методов машинного обучения
для решения задачи анализа биологических данных

Е. Б. Ахиезер, О. И. Дунаевская, И. В. Сердюк, С. В. Спивак

По статистике, каждая пятая супружеская пара сталкивается с невозможностью зачатия ребенка. Мужские половые клетки очень уязвимы, растущее число случаев мужского бесплодия подтверждает, что в современном мире очень много факторов, которые влияют и на активность сперматозоидов и на их количество. И важно не столько их количество, сколько качество. Спермограмма является объективным методом лабораторной диагностики, что позволяет максимально точно оценить способность к оплодотворению человека, проанализировав эякулят по ряду важнейших параметров. Только спермограмма способна ответить на вопрос о возможном мужском бесплодии и о наличии урологических заболеваний. При построении спермограммы, важно определять не только количество хороших сперматозоидов, но и их морфологию и подвижность. Поэтому исследования и совершенствования некоторых этапов спермограммы и является целью исследования. В данной статье решается задача классификации сперматозоидов на добрые и плохие, с учетом их подвижности и морфологии, с применением методов машинного обучения. Для реализации первого этапа машинного обучения (с учителем) в графическом редакторе были созданы учебные экземпляры (тренировочная выборка). Обучение было реализована тремя методами: методом опорных векторов, логистическая регрессия и метод К - ближайших соседей. По результатам тестирования выбран метод К - ближайших соседей. На этапе тестирования использовалась выборка из 15 различных сперматозоидов в различных вариациях вращения вокруг своей оси. Тестовая выборка не содержала экземпляров с тренировочной выборки и была сформирована с учетом морфологических особенностей сперматозоидов, но не копировала их с тренировочной выборки. На завершающем этапе обучения работе программы были протестированы на реальных данных.

Ключевые слова: машинное обучение; спермограмма; морфология; подвижность; распознавание образов; бинарная классификация; метод К-ближайших соседей.