FORMATION OF A DATABASE WITH DETERMINATION OF THE INFORMATIVENESS OF ITS CONTENT FOR THE DEVELOPMENT OF A PROGRAM FOR MULTIFACTOR PREDICTION OF THE OCCURRENCE OF BRUISING IN VOLUNTEER ATHLETES

Oleksandr Mokaniuk[1], Gunas Valery[1], Vitalii Mokin[2], Olena Slobodianiuk[2], Roman Kartelian[1], Artem Murenko[1]

[1]National *Pirogov* Memorial Medical University, Vinnytsia, Ukraine
[2]National Technical University, Vinnytsia, Ukraine

Summary: *Introduction. The development of forensic medicine in the era of modern information technologies requires the search for the most informative and accessible diagnostic methods. Materials and methods. The study was conducted on 22 volunteers (athletes). In them, 43 bruises were described, which were localized on the shoulder, trunk and thighs. Each bruise was photographed at the same time (23 time intervals) and described according to a standard scheme. Their localization, size, type of color and the presence of skin edema in the area of damage were determined. Results. At the first stage of work, all results were compiled into one table (data bank). A total of 795 bruises were described. An exploratory analysis of the data was then carried out, as a result of which all bruises that were less than an hour old were reduced to a whole numerical value (0 hours). The number of observations was reduced to 20. At the third stage, a comprehensive assessment of the correlation of diagnostic criteria for the age of occurrence of the remaining 707 bruises was carried out. It was established that the age of the bruise is reliably correlated with its color, and has a reliable negative correlation with its swelling. There is also a slight dependence on the age, sex of athletes, localization of injuries. At the next stages of the analysis, the hours of examination of bruises were transferred into 20 time intervals and their agglomerative clustering was carried out. Its purpose was to check the validity of these intervals. Which led to a reduction in their number from 20 to 15, within which the studied diagnostic criteria are not mixed. Conclusions. The most informative characteristics of a bruise are its color and the presence or absence of skin swelling. The obtained data are sufficient for the development of an experimental model of multifactor prediction of the age of bruises using neural networks. To increase the accuracy of the diagnosis, it is worth continuing the study to increase the number of observations.*

Keywords: *bruise, duration of bruising, multifactorial prediction.*

Introduction: Forensic medical examination of victims in cases of domestic and gender violence is very complex due to the fact that it covers a large volume of various issues, one of them is, in particular, determining the age of skin damage. The development of forensic medicine in the era of modern information technologies requires the search for the most informative, rational and accessible methods of forensic diagnostics, which should lead to an increase in their effectiveness. In the future, this will allow new expert diagnostic programs to be introduced into forensic medical practice [1]

The purpose of the study: Based on the study of bruises that occur in athletes while playing paintball, to form a database that includes the characteristics of a person and injuries, with the determination of the informativeness of its content for the development of a program for predicting the antiquity of their occurrence based on neural network algorithms.

Materials and methods: The study was conducted on 22 volunteers (athletes) who had been playing paintball for at least a year and knew the rules of the game and safety techniques. Each of the volunteers was familiarized with the rules of conducting the research and a contract was concluded with each of them (informed voluntary consent to conduct research with human participation). Among them were 8 women and 14 men aged 19 to 27. The participants of the study are physically healthy, did not apply medicinal products to the skin and did not use medications during the entire period of the experiment.

The game used a paintball marker that uses working gas (compressed air) to shoot paintballs (diameter 18 mm and weight 3.2 grams) with a flight speed of 150-300 feet per second. For the safety of the participants, a protective mask and suit were used, the textile material of which ensured maximum contact of the projectile with the skin.

During the examination of volunteer athletes, 43 bruises were described in them, which were localized on the shoulder, trunk and hips. Each bruise was photographed together with a color ruler with a metric scale located in the same plane as the damage, according to a technique that was developed earlier [2]. Photographs were taken 30 seconds after occurrence, 5 min., 15 min., 30 min., 1 h., 3 h., 6 h., 9 h., 12 h., 24 h., 48 h., 72 h., 96 h., 120 h., 144 h., 168 h., 192., 216 h., 240 h., 264 h., 288 h., 312 h., 336 h. A total of 23 research terms.

The research used a NikonAF-SNikkor 18-55 mm digital camera, the distance from the camera to the bruise was within 40 cm, the camera was located perpendicular to the bruise, and the lighting lamp was at an angle of 45°, ensuring the illumination of the research object at 1000-2000 Lux.

Bruises were described according to a standard scheme, their localization was determined (second group - shoulder and upper arm; third - trunk; fourth - thigh and buttocks), size (area in cm2), type of color (6 types: First type - red or purple; second type - red and purple, or blue-red; the third type - red and yellow-green or yellow; the fourth type - purple (red-violet, blue-red); the fifth type - magenta (red-violet, blue- red) and yellow-green, or yellow; the sixth type - yellow-green, or yellow) and the presence of skin edema in the area of damage (presence/absence).

A laptop computer ASUS Vivobook 15 X1500EA-BQ3733 was used for data processing.

Results: At the first stage of the work, all the results were summarized in one table - a data bank (Table 1), in which the serial number of the bruise, the serial number of the volunteer, the age of the victim - the number of full years, gender - male (1) or female (2) were indicated. duration of damage in hours, localization (6 types), color type (6 options), presence of skin edema (1), its absence (0). A total of 795 bruises were described. Data on the area of the bruise were not entered due to the fact that they were of similar size.

At the second stage of the study, an exploratory data analysis was carried out, namely, the data in the table were checked and all bruises that were less than an hour old were reduced to a whole numerical value (0 hours), accordingly, 88 rows were deleted. Number of observations reduced from 23 to 20.

At the third stage of the work, a comprehensive assessment of the correlation of diagnostic criteria for the age of occurrence of the remaining 707 bruises was carried out. The correlation coefficient r was determined, with a basic reliability of 99.9%. The obtained results are presented in Table 2.

At the fourth stage of the analysis, specific hours of bruise research were divided into 20 time intervals (classes), followed by their agglomerative clustering. Its purpose was to check the validity of these intervals, namely whether the studied diagnostic criteria do not intersect within these limits.

For ease of evaluation, the results were visualized using the tSNE method (Figure 1). Each cluster has its own color. [3,4,5,6].

As can be seen in Figure 1, the clusters are mixed, that is, a series of time intervals whose diagnostic criteria overlap should be combined.

This led to the need at the next stage to use a typical set of techniques for working with features - "feature engineering" and filtering anomalies (time intervals where the characteristics of bruises intersect). The first method allows you to identify the marginal values of the diagnostic criteria, where the volume of informative data is the smallest, and the second - to filter these values accordingly, which is one of the ways to improve the accuracy of the diagnosis. Data analysis was carried out by quartiles: 5%, 10%, 90% and 95%.

The results of the analysis are given in Table 3.

The analysis shows that less than 10% of the data characterizing the object of the study is the athlete's age is less than 19 years, and if we evaluate the bruises themselves, then 90% of the injuries are in the time range of up to 240 hours. Therefore, it is advisable to apply the following data filter:

data = data[(data[`age`] > 19) & (data[`damage_during`] <= 240)]

Applying this filter allows you to remove 96 injuries that belong to athletes younger than 19 years, or have a post-traumatic period of more than 240 hours, and the remaining bruises (611) are better equipped with data than 707.

As a result of these actions, the number of clusters in the study decreased from 20 to 16.

At the next stage, agglomerative clustering of the newly obtained data was again carried out (16 clusters – 611 bruises). The results were visualized using the tSNE method (Figure 2). Each cluster has its own color.

As can be seen in Figure 2, the brown and green clusters are still mixed, that is, these 2 intervals are not valid. Therefore, let's reduce the number of clusters from 16 to 15 (Figure 3):

As can be seen from Figure 3 for 15 clusters, the separation of data is more clear and, accordingly, there is no intersection of diagnostic criteria.

Discussion: To develop a program for multifactorial prediction of the age of bruises, it is necessary to first create a database with a determination of the informativeness of its content. In our study, the database is formed from 7 diagnostic criteria, 2 of which characterize a person, and 5 - an injury that occurred to him.

It was established that the age of the bruise reliably (p= 0.999) correlates with its color (r= 0.7728), and has a reliable (p= 0.999) negative correlation with its swelling (r= -0.5879), i.e. the older the injury bruising, the less likely it is to have skin swelling. There is also a slight dependence between other factors: age, gender of athletes, localization of injuries. The size of the damage in our study was not taken into account, since they were of the same type.

After assessing the informativeness of the diagnostic criteria, agglomerative clustering of the investigated time intervals was carried out, which made it possible to identify (invalid) clusters where the data are mixed. And a typical set of techniques for working with "feature engineering" signs and filtering anomalies is to remove these intervals, reducing their number from 20 to 15.

The database formed in this way in the defined 15 time intervals can be used in the future to develop programs for multifactorial prediction of the age of the bruise using neural networks.

Conclusions: 1) The most informative characteristics of a bruise for determining the age of the injury are the color of the bruise and the presence or absence of skin edema.

Due to the fact that all bruises were caused by the same paintballs and had similar sizes, determining the effect of such a characteristic of the bruise as its size requires further research

2) The obtained data are sufficient for the development of an experimental model of multifactorial prediction of the age of bruises, under the condition of reducing the number of time intervals from 20 to 15;

3) It is worth continuing the research to increase the number of observations, which will allow to increase the accuracy of the diagnosis of the age of the injury.

Bibliographic images:

Table 1. Combined table of diagnostic criteria for the age of bruising.

Serial number of the bruise	Volunteer number	Age	Sex	Age of injuries	Localization	Color type	Edema
1	1	20	2	0,003	2	1	0
2	1	20	2	0,05	2	1	0
3	1	20	2	0,15	2	1	0
4	1	20	2	0,3	2	1	1
5	1	20	2	1	2	1	1
6	1	20	2	3	2	1	1
7	1	20	2	6	2	1	0
8	1	20	2	12	2	2	0
9	1	20	2	24	2	4	0
10	1	20	2	48	2	4	0
***	***	***	***	***	***	***	***
795	9,2	20	2	264,000	3	6	0

Table 2. The results of the statistical analysis of the correlation of the diagnostic criteria for the age of bruising

	index	sex	age	damage_during	localization	color	edema
index	1.0000	0.7909	0.2114	0.0389	0.5841	0.0111	0.0711
sex	0.7909	1.0000	0.3573	-0.0981	0.1821	-0.0014	-0.0055
age	0.2114	0.3573	1.0000	0.0971	-0.1723	-0.0279	0.0114
damage_ during	-0.0389	-0.0981	-0.0971	1.0000	0.0257	0.7728	-0.5879
localization	0.5841	0.1821	-0.1723	0.0257	1.0000	-0.0121	0.0864
color	0.0111	-0.0014	-0.0279	0.7728	-0.0121	1.0000	-0.6063
edema	0.0711	-0.0055	0.0114	-0.5879	0.0864	-0.6063	1.0000

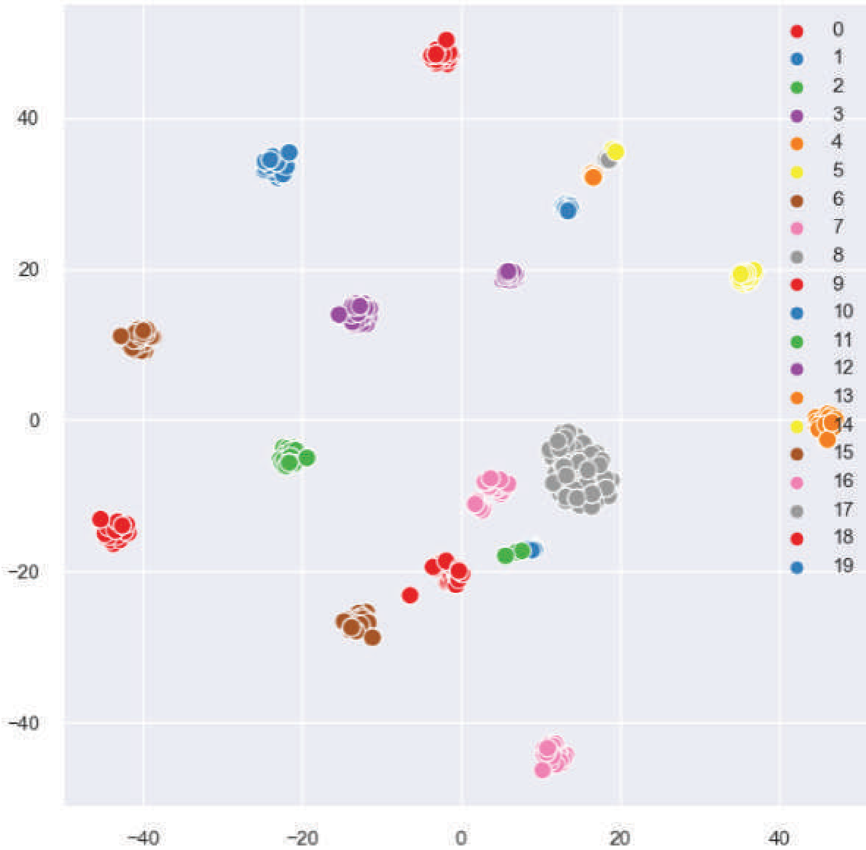

Fig. 1. Visualization of the results of agglomerative clustering (20 clusters) by the tSNE method

Table 3. The results of the processing of "feature engineering" diagnostic features from the database.

	sex	age	damage_during	localization	color	edema
count	707.000000	707.000000	707.000000	707.000000	707.000000	707.000000
mean	0.509194	20.606789	96.782178	3.115983	4.294201	0.288543
std	0.500269	1.430489	97.504242	0.843648	1.813862	0.453406
min	0.000000	19.000000	0.000000	2.000000	1.000000	0.000000
5%	0.000000	19.000000	0.000000	2.000000	1.000000	0.000000
10%	0.000000	20.000000	0.000000	2.000000	1.000000	0.000000
25%	0.000000	20.000000	6.000000	2.000000	4.000000	0.000000
50%	1.000000	20.000000	72.000000	3.000000	5.000000	0.000000
75%	1.000000	21.000000	168.000000	4.000000	6.000000	1.000000
90%	1.000000	21.000000	240.000000	4.000000	6.000000	1.000000
95%	1.000000	25.000000	288.000000	4.000000	6.000000	1.000000
max	1.000000	25.000000	336.000000	4.000000	6.000000	1.000000

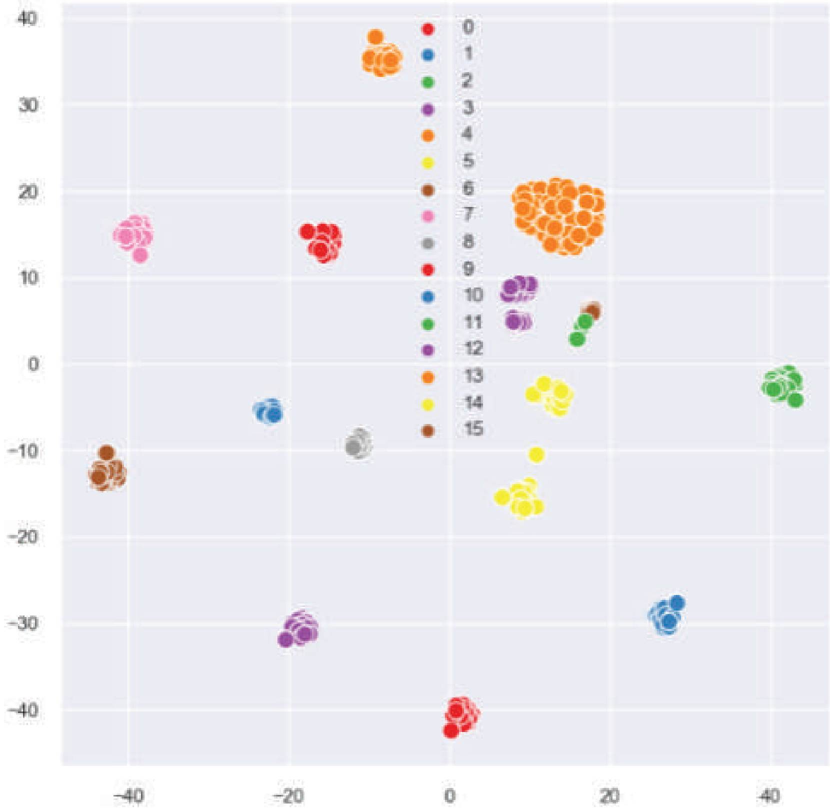

Fig 2. Visualization of the results of agglomerative clustering (16 clusters) by the tSNE method

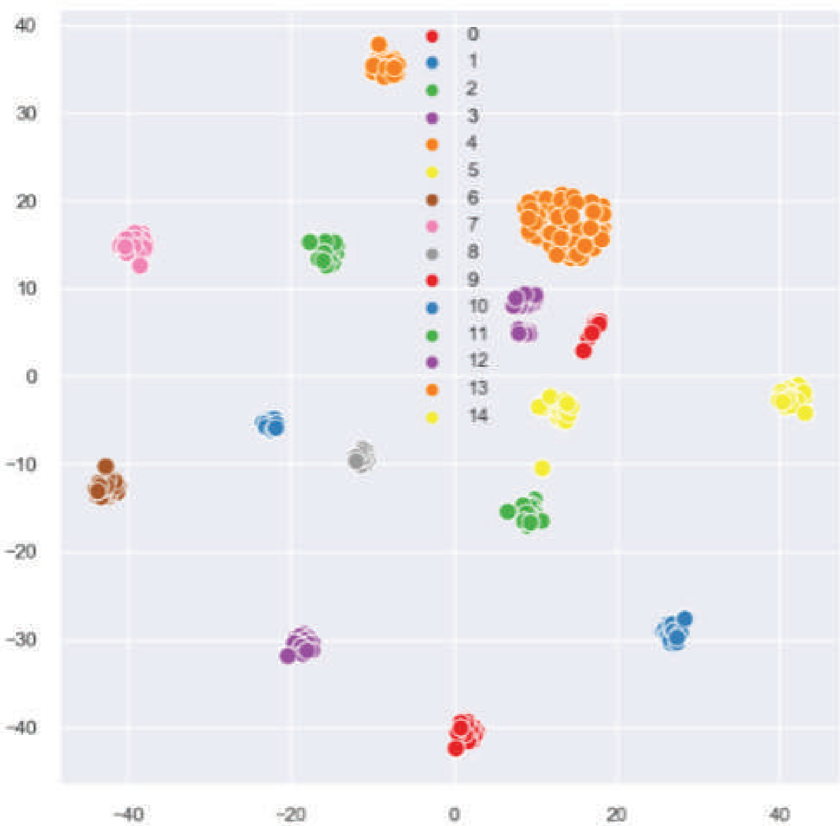

Fig 3. Visualization of the results of agglomerative clustering (15 clusters) by the tSNE method

References:

1. Mokaniuk OI, Havryliuk AO, Perebetiuk AM, Dzhurabaiev VS, Olchedai VM. Vstanovlennia koloru travmovanoi shkiry za dopomohoiu shkaly koloriv. Sudovo-medychna ekspertyza. 2011(4):15-16. (in Ukrainian).
2. Mokaniuk OI. Suchasnyi pidkhid shchodo vyznachennia koloru syntsia pry sudovo-medychnii ekspertyzi zhyvykh osib (analiz arkhivnoho sudovo-medychnoho materialu). Zbirnyk naukovykh prats spivrobitnykiv KMAPO im. P.L. Shupyka. 2001 (10):1310-1317. (in Ukrainian).
3. L.J.P. van der Maaten, G.E. Hinton. Visualizing Non-Metric Similarities in Multiple Maps. Machine Learning 87(1):33-55, 2012.
4. L.J.P. van der Maaten. Learning a Parametric Embedding by Preserving Local Structure. In Proceedings of the Twelfth International Conference on Artificial Intelligence & Statistics (AI-STATS), JMLR W&CP 5:384-391, 2009.
5. L.J.P. van der Maaten. Accelerating t-SNE using Tree-Based Algorithms. Journal of Machine Learning Research 15(Oct):3221-3245, 2014.
6. Belkina, A.C., Ciccolella, C.O., Anno, R. et al. Automated optimized parameters for T-distributed stochastic neighbor embedding improve visualization and analysis of large datasets. Nat Commun 10, 5415, 2019.
7. Mokanyuk, O., Perebetiuk, A., & Gunas, V. (2019). Determination of the informativeness of the criteria for diagnosing timing of bruising by the Cox model. Folia Societatis Medicinae Legalis Slovacae 9 (1), 28-30
8. Mokanyuk, O., Perebetiuk, A., & Gunas, V. (2019). Logic-mathematical analysis of the informativeness of the criteria for diagnosing timing of bruising. 100 years of institute of forensic medicine 1919-2019, 96-101
9. Gunas, V., Bobkov, P., Plakhotniuk, I., Olhovenko, S., & Solonyi, O. (2021). Specifics of fire damage to cotton clothing while shooting point-blank at a human torso simulator from a Fort-12RM pistol. Theory and Practice of Forensic Science and Criminalistics, 23(1), 175-187.

Mokaniuk Oleksandr, expert of forensic medicine, PhD, Docent of Forensic Medicine and Law Department National Pirogov Memorial Medical University, Vinnytsia, Ukraine,
e-mail: mokaniuk1964@gmail.com, ORCID ID: 0000-0003-4544-4441.