Proceedings of the ASME 2019 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference
IDET/CIE 2019
August 18-21, 2019, Anaheim, CA, USA

IDETC2019-97902

AUGMENTED MICROSCOPY FOR DNA DAMAGE QUANTIFICATION: A MACHINE LEARNING TOOL FOR ENVIRONMENTAL, MEDICAL AND HEALTH SCIENCES

Michele Bernardini*, Alessandro Ferri, Lucia Migliorelli, Sara Moccia, Luca Romeo
VRAI Laboratory,
Department of Information Engineering
Università Politecnica delle Marche
Brecce Bianche 60131, Ancona
{m.bernardini, a.ferri, l.migliorelli}@pm.univpm.it
s.moccia@staff.univpm.it, l.romeo@univpm.it

Sonia Silvestri, Luca Tiano
Department of Life and
Environmental Sciences (DISVA)
Università Politecnica delle Marche
Brecce Bianche 60131, Ancona
l.tiano@univpm.it,
s.silvestri@univpm.it

Adriano Mancini
VRAI Laboratory,
Department of Information Engineering
Università Politecnica delle Marche
Brecce Bianche 60131, Ancona
a.mancini@univpm.it

ABSTRACT

The Comet Assay is a well-known procedure employed to investigate the DNA damage and can be applied to several research areas such as environmental, medical and health sciences. User dependency and computation time effort represent some of the major drawbacks of the Comet Assay. Starting from this motivation, we applied a Machine Learning (ML) tool for discriminating DNA damage using a standard hand-crafted feature set. The experimental results demonstrate how the ML tool is able to objectively replicate human experts scoring (accuracy detection up to 92%) by solving the related binary task (i.e., controls vs damaged comets).

Introduction

DNA is the repository of genetic information in cells therefore its integrity and stability are essential to life. However it is not inert but subject to assault from the environment, and any resulting damage could lead to mutation and possibly disease. Perhaps the best-known example of the link between environmental-induced DNA damage and disease is skin cancer, caused by excessive exposure to UV radiation present in sunlight. Another example is the damage caused by tobacco smoke, which can lead to mutations and subsequent lung cancer. Besides environmental agents, DNA is also subject to oxidative damage from products of metabolism, such as free radicals. In fact, an individual cell can suffer up to one million DNA lesions per day and their accumulation is one of the characteristics of the ageing process. Detection of DNA damage is therefore of paramount importance in different fields of basic and applied medical and health sciences, including environmental studies for verifying the toxicity of xenobiotics or chemicals released in the environment [1]. Among the different methodos developed to quantify DNA damage, the single cell electrophoresis or Comet Assay is prominent being a simple, rapid and sensitive method for measuring DNA breaks in clusters of cells [2]. The resulting image observed under the microscope appears as a "comet" with a distinct head and tail. The head is composed of intact DNA, while the tail consists of damaged (single-strand or double-strand breaks) or broken pieces of DNA. When standardized and validated, the Comet Assay can provide invaluable information in the areas of hazard identification and risk assessment of environmental and occupational exposure [1], diseases linked with oxidative stress (e.g., diabetes and cardiovascular disease) [3], nutrition [4], monitoring the effectiveness of medical treatment and investigating individual variation in response to DNA damage that may reflect genetic or environmental influences. The information obtained could lead to individual advice on lifestyle changes to promote health and especially on relative risks of genotoxic exposure to

*Address all correspondence to this author.

environmental pollution in humans, in sentinel organisms or in *in vitro* toxicity studies [5].

Despite its popularity, the Comet Assay still has some shortcomings mainly due to (i) a high inter-operator variability, (ii) the inter-laboratory variability and (iii) the high time effort. The advancement of research enabled the development of completely automated acquisition software/hardware systems, that combine operator-independent procedure and high-processivity capabilities. However, in the literature a lack of standard criteria unanimity accepted by the experts still remains [6, 7]. In fact, some ambiguities exist on which features the DNA damage should be discriminated, and moreover, the decision about the severity of the comet damage is rarely automatized, but still entrusted to human experts [8]. This condition can be no longer accepted when thousands of cells have to be analyzed and discriminated [7].

Since high-throughput of image acquisition is already allowed by a new generation of automated microscope readers, the aim of the current work is to present a novel approach of comet images classification based on automatized and reliable Machine Learning (ML) approach.

Specifically, the present work is finalized to provide improvement in the Comet Assay methodology through the combination of:

- Fully automated, operator-independent and high-throughput image acquisition and features extraction system;
- ML techniques in order to discriminate and quantify the DNA damage.

Current scenario

Comets can be extracted from the background and distinguished in head and length by expert's visual inspection or image analysis software. Even if an accurate visual scoring inspection made by and expert can provide an immediate qualitative indication of severity of the DNA damage, the time employed by experts to score the images represents the major drawback of the Comet Assay. Moreover, the inspection is totally subjective and cannot be reproduced, standardized or compared by other researchers. On the contrary, the employment of comet analysis tools lead to overcome only a part of these issues by extracting some salient features. Moreover, some of these software still suffer from user intervention by setting indispensable threshold parameters, while fully-automated systems are completely user independent. Nowadays, publicly available automated software are very popular [9–11]. OpenComet [9] is a open-source, scalable and useful tool, but still is not able to only extract hand-crafted features without providing the classification of the damage.

Method

Computer-based image analysis provides an objective method of scoring visual content independent of subjective manual interpretation, while potentially being more sensitive, consistent and accurate. The computer system automatically assigns images to user-defined image classes extrapolated from an experimental control. Extraction of numerical descriptors from reference experimental dataset of images enabled classification using ML tools of test images. The training data were used to automatically define the classification rules, while the test data were used to assess the effectiveness of these rules and their ability to consistently reflect the data. This enables the development of a ML tool able to explore and elaborate image data and train the model, followed by reliable real-time predictions.

Comet assay procedure

The comet assay was carried out to measure DNA damage as previously described [12]. Briefly, Aliquots of γ rays irradiated cells (0-450 cGy) and control cells containing 10 000 cells from each sample were transferred to eppendorf tubes and centrifuged for 10 min at 800 g at 4° C. The supernatant was removed and the cells were resuspended in 0.7% low melting agarose from which 0.035 ml were taken and placed on pre-coated, high throughput, comet assay slides (Trevigen). Trevigen Comet hightrouput microscope slides are characterized by clean area separated by silicon barriers in order to allow simultaneous layering of ten different samples on each slide. Clean areas are manifactured with a dried agarose coating in order to enhance adhesivity. The microgels on the slides were then allowed to solidify at 4°. Subsequently, the slides were immersed overnight at 4° in the dark, in ice-cold, freshly prepared lysis solution (2.5 M NaCl, 100 mM Na2EDTA, 10 mMTris-HCl, 1% Triton X-100 and 10% DMSO, adjusted to pH 10) in order to lyse the embedded cells and to allow DNA unfolding. After incubation in lysis solution, the slides were exposed to alkaline buffer (1 mM Na2EDTA, 300 mM NaOH buffer, pH 13 for 30 min to allow DNA unwinding. Electrophoresis was then performed for 20 min at 1 V/cm in the same buffer. After neutralization in Tris buffer (pH 7.5) and dehydration in 75% methanol, the DNA on each slide was stained with 0.015 ml ethidium bromide (20μg/ml) and viewed under fluorescent light using an Olympus BX51 fluorescence microscope.

Data and Features extraction

For each sample, 15 randomly acquired images were recorded and processed using a custom made software Marche [13] based on Labview programming platform (National Instruments) that enables automatic identification of the comets, greatly reducing operator-dependent variability. A key feature of the software is its ability to identify the comets from the background and to estimate the commonly used DNA-damage indexes. Comet specific DNA-damage indexes and images of 150 nucleoids for each slide were fed to a database. Three slides were analyzed for each treatment condition and a total of 335 comets

(a) Subset of control comets

(b) Subset of damaged comets

FIGURE 1: Control and damaged comets: region of interest (ROI) and the centre of the head are evidenced in red for each comet.

was selected (the other comets are not included in our analysis). Figure 1 shows an example of a subset of the collected dataset. Three expert biologists labeled the whole dataset resulting of 256 control comets (Fig. 1a) and 79 damaged comets (Fig. 1b). The final label of the image was provided according to a majority vote approach.

Digital images have been appropriately calibrated [14] in order to associate geometrical units (μm) with pixels at the acquired magnification (20x). Each pixel of the 8bit image has a grayscale index varying from 0 (black) to 256 (white) proportional to the fluorescence intensity of the DNA-bound probe, that is ultimately directly proportional to the amount of DNA. The sum of pixel intensities along the Y axis in the region of interest (ROI) defining the comet results in a histogram characterized by three major values on the X axis that are the beginning and the end of the ROI and the value corresponding to the peak of intensity. This can be single in case of undamaged or lightly damaged cells or may be multiple in case of heavy damaged cells. In the latter case only the first peak will be considered. The area comprised from the beginning of the ROI (*Comet Start*) and the first peak in the intensity profile (*CometPeak1*) defines half of the comet head typically composed by intact DNA. In intact comets, the comet head equals the total area of the comet (i.e., all the DNA is intact and the damaged DNA, also identified as comet tail is null and as a result the histogram profile is a Gaussian curve). In damaged comets total area is greater than comet head and image analysis can be applied in order to calculate useful DNA damage indexes, both geometric and light intensity parameters. The nine indexes used in this study are detailed below:

1. *Head length* (μm) is derived from number of pixels in horizontal direction of the comet and represents the length corresponding to $2 * (\textit{CometPeak1} - \textit{Comet Start})$.
2. *Head intensity* (%) is defined as (sum of all pixel intensity values in the comet head / sum of total intensities in the ROI) $*100$. It is proportional to the percentage of intact DNA.
3. *Tail length* (μm) is defined as the *Head length* subtracted from the overall comet length. In damaged comets it exceeds the value of 1/2 *Head length*, but is subjected to saturation not showing a linear response over a wide range of DNA damage.
4. *Tail intensity* (%) is defined as $100 - \textit{Head intensity}$.
5. *Tail moment* is defined as the percentage of DNA in the comet tail multiplied by the *tail length*.
6. *Tail migration* is proportional to the *tail length*.
7. *Total area* is defined as the number of pixels enclosed in the comet shape.
8. *Gray mean level* is defined as the average of the grey levels from pixels in the comet.
9. *Total intensity* is defined as the sum of all pixel intensity values in the comet.

All the features were used as the predictors of the ML models, while the label is the absence or presence of damage.

Statistical analysis

The statistical analysis aimed to evaluate and quantify the dependency between the extracted predictors and the comet damage. In particular, we tested the null hypothesis that the data (features set) comes from a normal distribution according to a one-sample Kolmogorov-Smirnov test at the 5% of significance level. Then, we employed the two-sample Kolmogorov-Smirnov test at the 5% of significance level in order to verify if there is a sort of dependency between the features observations and the related label (controls vs damaged comets). This test measures the distance between the empirical distribution functions of two samples (i.e., predictors versus labels to be predicted).

Machine Learning analysis and Measures

The ML tool aims to estimate the binary label of the comets through the predictors extracted by the automated image processing. Specifically, the following standard ML models were employed and compared:

– Decision Tree (DT) [15]

– Random Forest (RF) [16]
– K-Nearest Neighbor (KNN) [17]
– KNN with NN features selection (KNN+NCFS) [18]
– Linear Support Vector Machine (SVM Lin) [19]
– Gaussian Support Vector Machine (SVM Gauss) [19]

The performance of the introduced ML models was evaluated according to the following measures:

– *Accuracy*: the percentage of correct predictions;
– *Macro-precision*: the percentage of true positive over the predicted condition positive. We refer to this metric with *Precision*. The *Precision* is calculated for each class and then take the unweighted mean;
– *Macro-recall*: the percentage of true positive over the condition positive (sensitivity). We refer to this metric with *Recall*. The *Recall* is calculated for each class and then take the unweighted mean;
– *Macro-F1*: the harmonic mean of *precision* and *recall* averaged over all classes. We refer to this metric with *F1*.

We computed in all the experiments a stratified Tenfold Cross-Validation (10-CV) over comets procedure. The hyperparameters optimization was performed implementing a grid-search and optimizing the *macro-recall* score in a nested stratified Five fold Cross-Validation. Hence, each split of the outer CV was trained with the optimal hyperparameters tuned in the inner CV. Despite the high computational cost of this procedure, it allows to obtain an unbiased and robust model checking evaluation. Table 1 shows for each ML the different hyperparameters as well as the grid-search range.

TABLE 1: Range of Hyperparameters (Hyp) for each model: Decision Tree (DT), Regression Forest (RF), K-Nearest Neighbor (KNN), KNN with NN features selection (NCFS), Linear Support Vector Machine (SVM Lin) and Gaussian Support Vector Machine (SVM Gauss).

Model	**Hyp**	**Range**
DT	max n° of splits	$\{5, 10, 15, 20, 25\}$
	min n° of leaf size	$\{50, 60, 70, 80, 90, 100\}$
RF	n° of DT	$\{50, 100, 150, 200, 250\}$
	n° of predictors to select	$\{\frac{all}{4}, \frac{all}{3}, \frac{all}{2}, all\}$
KNN	n° of neighbors	$\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$
KNN+NCFS	Regularization parameter λ	$\{10^{-4}, 10^{-3}, 10^{-2}, 0.1, 1\}$
SVM Lin	Box Constraint	$\{10^{-3}, 10^{-2}, 0.1, 1, 10\}$
SVM Gauss	Box Constraint	$\{10^{-2}, 0.1, 1, 10, 10^2, 10^3, 10^4\}$
	Kernel Scale	$\{10^{-2}, 0.1, 1, 10, 10^2, 10^3, 10^4\}$

Results

Statistical analysis

The performed one-sample Kolmogorov-Smirnov test rejects the null hypothesis ($p < 0.05$) that the data (features set) comes from a normal distribution. Accordingly, the two-sample Kolmogorov-Smirnov test demonstrates how the nine extracted features have a statistically significant ($p < 0.05$) dependency with respect to the related label. These results confirm the effectiveness of the employment of these features as predictors of the ML model in order to discriminate the damaged comets. Figure 2 shows the histogram plot for each different feature and for each condition (i.e., control vs damaged comet). The histograms confirm qualitatively the discriminative power of the extracted features in order to perform the classification task.

Machine Learning approaches

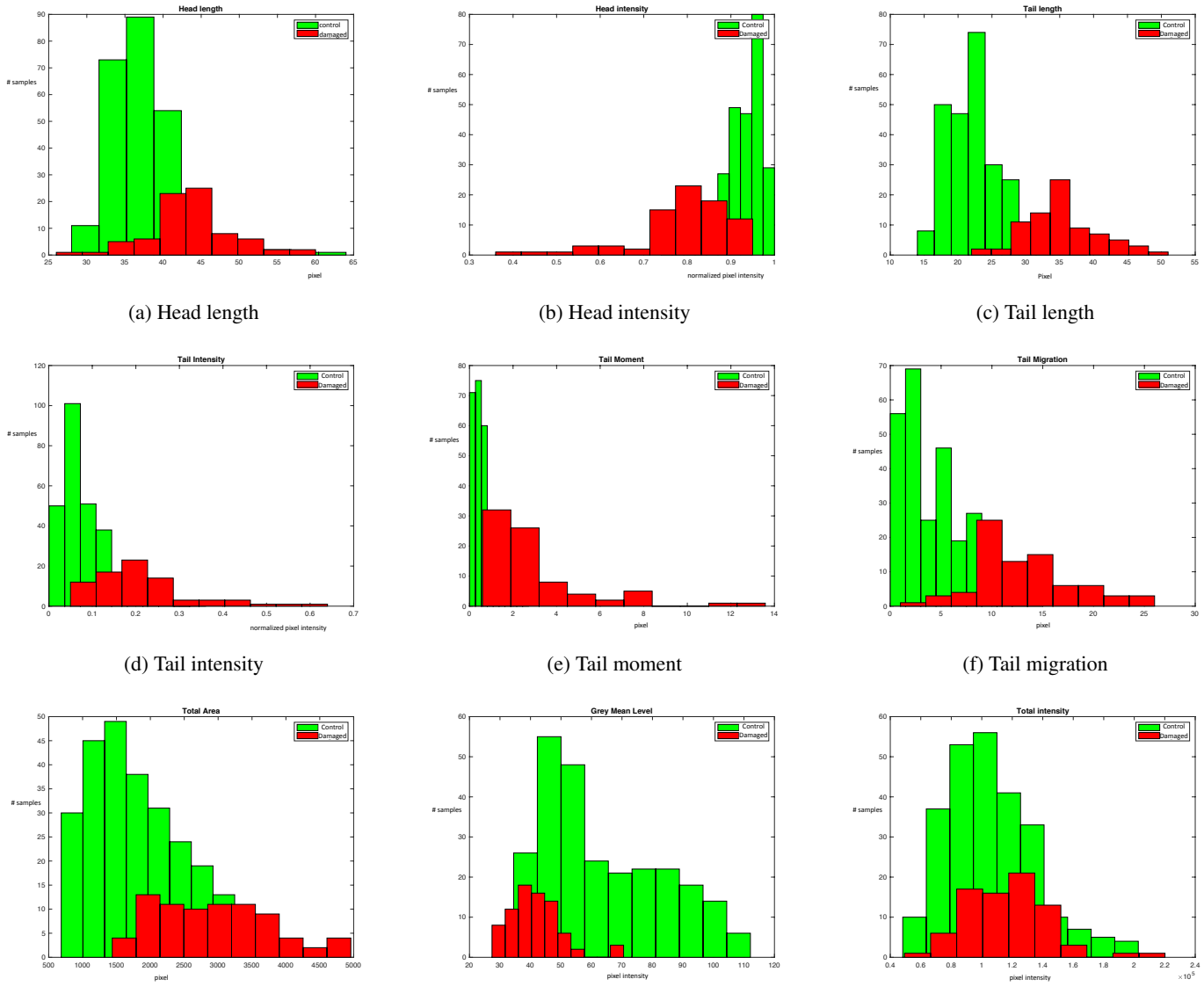
Table 2 shows the results of the proposed ML models for discriminating the damage of comets. The SVM Lin achieved the best performance in terms of *Accuracy* (0.92 ± 0.03), *F1* (0.88 ± 0.05), *Precision* (0.89 ± 0.05) and *Recall* (0.89 ± 0.06). The performance is stable across the 10-CV fold providing a minimum value of 0.89, 0.80, 0.84 and 0.77, respectively.

TABLE 2: Machine Learning results: Decision Tree (DT), Regression Forest (RF), K-Nearest Neighbor (KNN), KNN with NN features selection (NCFS), Linear Support Vector Machine (SVM Lin) and Gaussian Support Vector Machine (SVM Gauss). The best results were highlighted in bold.

Model	*Accuracy*	*F1*	*Precision*	*Recall*
DT	0.88	0.83	0.84	0.83
RF	0.90	0.86	0.87	0.86
KNN	0.76	0.66	0.67	0.66
KNN+NCFS	0.86	0.80	0.80	0.79
SVM Lin	**0.92**	**0.88**	**0.89**	**0.89**
SVM Gauss	0.89	0.84	0.85	0.83

Discussion

In this work we set up a framework that consists of a fully automated and high-throughput image acquisition and features extraction system combined with a ML tool, in order to discriminate the presence or the absence of the DNA damage. We proposed a reliable ML tool able to replicate manual scoring with an accuracy detection up to 92%. The main benefits of applying ML techniques for solving this binary task (i.e., controls vs



(a) Head length (b) Head intensity (c) Tail length

(d) Tail intensity (e) Tail moment (f) Tail migration

(g) Total area (h) Total gray level (i) Total intensity

FIGURE 2: Statistical analysis.

damaged) consist to avoid the subjective bias and to exponentially decrease the time-consuming process. The application of the methodology is interdisciplinary, and can range in macro-areas such as cancer research, ecotoxicology and skin health. Thus, the improved methodology can enable rapid, automatic, operator-independent, measurement of DNA damage in relation to different health-related issues. More specifically, the proposed analytical platform can also likely foster its use in other research fields such as diseases linked with oxidative stress (e.g., diabetes and cardiovascular disease), human and animal fertility, ageing research, nutrition and sports medicine. Basically, the augmented microscopy can enhance the effectiveness of medical treatment and the monitoring of individual variation in response to DNA damage that may reflect genetic or environmental influences.

Conclusions and future work

Since the proposed method still represents a preliminary approach, a future development would be to extend the binary problem to a multi-class problem, where the damage class is divided at least in other three intermediate sub-classes. In this future direction, taking into account the higher difficulty of the task, ML

and Deep Learning (DL) approach can be combined in order to (i) segment the comet (i.e., using a CNN as made in the very similar problem of head segmentation in top-view images [20, 21]), (ii) discover a novel highly discriminative feature descriptors (without using hand-crafted features) and (iii) learn some hidden patterns which can be sometimes unsighted by biologists. This aspect can increase the usefulness of the framework for solving increasingly important clinical challenges.

Additionally, it would be of considerable interest to investigate the DNA damage identification with ML unsupervised approaches (e.g., spectral clustering, hierarchical clustering, Gaussian mixture model) in order to minimize as much as possible the human labeling procedure.

Considering the processing potentiality of the framework, we plan to increase the number of comets included in the dataset (up to 5k comets). In this scenario, the ML tool can be integrated in a cloud-framework and can be updated continuously, because all the input and output data of the algorithm can be shared among researchers.

ACKNOWLEDGMENT

This work was supported in part by the University strategic project "Augmented microscopy for DNA damage quantification: a key tool for environmental, medical and health sciences".

REFERENCES

[1] Costa, P. M., Pinto, M., Vicente, A. M., Gonçalves, C., Rodrigo, A. P., Louro, H., Costa, M. H., Caeiro, S., and Silva, M. J., 2014. "An integrative assessment to determine the genotoxic hazard of estuarine sediments: combining cell and whole-organism responses". *Frontiers in Genetics*.

[2] Fairbairn, D. W., Olive, P. L., and O'Neill, K. L., 1995. "The comet assay: a comprehensive review". *Mutation Research/Reviews in Genetic Toxicology,* ***339***(1), pp. 37–59.

[3] Takahashi, M., Keicho, K., Takahashi, H., Ogawa, H., Schulte, R., and Okano, A., 2000. "Effect of oxidative stress on development and dna damage in in-vitro cultured bovine embryos by comet assay". *Theriogenology,* ***54***(1).

[4] Wasson, G. R., McKelvey-Martin, V. J., and Downes, C. S., 2008. "The use of the comet assay in the study of human nutrition and cancer". *Mutagenesis,* ***23***(3), pp. 153–162.

[5] Collins, A. R., 2004. "The comet assay for dna damage and repair". *Molecular Biotechnology,* ***26***(3), p. 249.

[6] Lovell, D. P., and Omori, T., 2008. "Statistical issues in the use of the comet assay". *Mutagenesis,* ***23***(3), pp. 171–182.

[7] Collins, A. R., El Yamani, N., Lorenzo, Y., Shaposhnikov, S., Brunborg, G., and Azqueta, A., 2014. "Controlling variation in the comet assay". *Frontiers in Genetics,* ***5***, p. 359.

[8] Vojnovic, B., Barber, P., Johnston, P., Gregory, H., Marples, B., Joiner, M., and Locke, R., 2013. "A high sensitivity, high throughput, automated single-cell gel electrophoresis (comet) dna damage assay". *Physics in Medicine & Biology,* ***58***(1), p. 15.

[9] Gyori, B. M., Venkatachalam, G., Thiagarajan, P., Hsu, D., and Clement, M.-V., 2014. "Opencomet: an automated tool for comet assay image analysis". *Redox Biology*.

[10] Końca, K., Lankoff, A., Banasik, A., Lisowska, H., Kuszewski, T., Góźdź, S., Koza, Z., and Wojcik, A., 2003. "A cross-platform public domain pc image-analysis program for the comet assay". *Mutation Research/Genetic Toxicology and Environmental Mutagenesis,* ***534***(1-2).

[11] "CometScore Comet Scoring Software, Version 2.0".

[12] Orlando, P., Silvestri, S., Galeazzi, R., Antonicelli, R., Marcheggiani, F., Cirilli, I., Bacchetti, T., and Tiano, L., 2018. "Effect of ubiquinol supplementation on biochemical and oxidative stress indexes after intense exercise in young athletes". *Redox Report,* ***23***(1), pp. 136–145.

[13] Tiano, L., Littarru, G. P., Principi, F., Orlandi, M., Santoro, L., Carnevali, P., and Gabrielli, O., 2005. "Assessment of dna damage in down syndrome patients by means of a new, optimised single cell gel electrophoresis technique". *Biofactors,* ***25***(1-4), pp. 187–195.

[14] Innocenti, B., Lambert, P., Larrieu, J.-C., Pianigiani, S., Paolanti, M., Bernardini, M., Cenci, A., and Frontoni, E., 2016. "Development of an automatic procedure to mechanically characterize soft tissue materials". In 2016 12th IEEE/ASME International Conference on Mechatronic and Embedded Systems and Applications (MESA), pp. 1–6.

[15] Breiman, L., Friedman, J., Olshen, R., and Stone, C., 1984. *Classification and Regression Trees*. Wadsworth and Brooks.

[16] Breiman, L., 2001. "Random forests". *Machine Learning,* ***45***(1), pp. 5–32.

[17] Paolanti, M., Romeo, L., Liciotti, D., Pietrini, R., Cenci, A., Frontoni, E., and Zingaretti, P., 2018. "Person re-identification with rgb-d camera in top-view configuration through multiple nearest neighbor classifiers and neighborhood component features selection". *Sensors,* ***18***(10).

[18] Yang, W., Wang, K., and Zuo, W., 2012. "Neighborhood component feature selection for high-dimensional data". *Journal of Computers,* ***7***(1), p. 161.

[19] Cortes, C., and Vapnik, V., 1995. "Support-vector networks". *Machine Learning,* ***20***(3), pp. 273–297.

[20] Liciotti, D., Paolanti, M., Frontoni, E., Mancini, A., and Zingaretti, P., 2016. "Person re-identification dataset with rgb-d camera in a top-view configuration". In *Video Analytics. Face and Facial Expression Recognition and Audience Measurement*. Springer, pp. 1–11.

[21] Liciotti, D., Paolanti, M., Pietrini, R., Frontoni, E., and Zingaretti, P., 2018. "Convolutional networks for semantic heads segmentation using top-view depth data in crowded environment". In ICPR 2018, pp. 1384–1389.