Image Processing in Synthesis and Optimization of Active Vaccinal Components

Oana-Constantina Margin[1], Eva-H. Dulf [1], Teodora Mocan [2,3], Lucian Mocan[4]

[1] *Department of Automation, Faculty of Automation and Computer Science, Technical University of Cluj-Napoca, Str. Memorandumului 28, 400114 Cluj-Napoca, Romania (e-mail: oana.margin@student.utcluj.ro)*

[2] *Department of Physiology, Iuliu Haţieganu University of Medicine and Pharmacy, 400000 Cluj-Napoca, Romania*

[3] *Nanomedicine Department, Regional Institute of Gastroenterology and Hepatology, 400162 Cluj-Napoca, Romania*

[4] *Department of Surgery, 3-rd Surgery Clinic, Iuliu Hatieganu University of Medicine and Pharmacy, 400000 Cluj-Napoca, Romania*

Abstract: Worldwide, cancer is the second cause of death after heart diseases, being accountable for 10 million deaths per year. This study addresses adenocarcinoma, the main subject to multiple anticancer treatments, that are currently developing in medicine and pharmacy study trials. A new research for a therapeutic vaccine involves the study of gold nanoparticles impacting the immune response for annihilating cancer cells. The model is proposed to be implemented using Quantitative-Structure Activity Relationship (QSAR) techniques, specifically artificial neural networks in relation with fuzzy rules to combine the benefits of human perception with the automatic characteristic of neural nets. The inputs to the resulted ANFIS model are molecular features that must be selected for optimization, using antlion optimization algorithm, inspired recently from natural behaviors, the same way once ANNs were developed. A couple of molecular features are extracted and computed from hyperspectral images through image processing approaches like morphological transformations and watershed segmentation.

Keywords: QSAR, ALO, ANFIS, watershed segmentation, vaccine, cancer, image processing

1. INTRODUCTION

A current problem that concern the entire human race nowadays is the treatment and prevention of cancer. Cancer represents the term used for defining malignant disorders of a healthy tissue. Tumors are developed when normal, healthy cells are replaced or transformed into cancerous cells under the exposure to external factors like chemicals, infection or radiation. Besides the fact that every organism has a different immune response, the process of cell multiplication is fast and uncontrollable, which makes it hard to treat. Cancer can appear in any organ, but the most encountered types are adenocarcinomas, type of cancer that spreads to organs which produce fluids in the glands, i.e. lungs, colon and rectum, prostate, breasts (Nancy Moyer, 2019). According to World Health Organization (WHO) statistics, in 2020, the occurrence of new cancer diagnosis has reached 19.3 million cases, among which the leading types are adenocarcinomas. In the last years, the increase in cancer incidence and mortality was caused by the rising rate of ageing and the modern lifestyle, reducing life expectancy (Sung, et al., 2021).

The importance of the research is due to the lack of drug developments that can offer totally successful results in treatment of adenocarcinomas. Since computer aided methods support more and more medicine and pharmaceutical advances, therapeutic vaccines can be developed and also offer efficient results in treatment of cancer. An anticancer vaccine, like the one that must be further modelled in this study, is determining the immune system to create antigens for destroying cancerous cells.

Newly proposed drugs or treatments must always be rigorously tested, firstly to assess their potential effect and secondly to eliminate any kind of risk associated with unexpected reactions to the medication. Besides many other advantages like costs and ethical considerations, Quantitative-Structure Activity Relationship (QSAR) models achieve successfully all the requirements (Pradeep, et al., 2020). The main goal of the current research, is to study the relations between vaccine components and molecules structure and determine the model of the vaccine, further utilized to predict cell activity under the influence of applied treatment.

QSAR models have been popularized in recent years, when advantages of using computer tools in medicine were noticed. The core of the method is the characteristic of structural similarities in molecules that lead to similar biological activity. Hence, one important aspect before developing a QSAR model is the determination of chemical structure by a variety of molecular descriptors from which a selection will constitute the function for predicting activity (Peter, et al., 2019). Non-linear models are the most suitable for representing biochemical processes, such that machine learning techniques

are the best approaches to be studied and used. Machine learning techniques such as support vector machines (SVM) and k-nearest neighbors (kNN) (Tang, et al., 2009), artificial neural networks (ANN), sometimes combined with fuzzy logic in adaptive neuro-fuzzy inference systems (ANFIS) (Elaziz, et al., 2018) are all efficient methods among the most employed non-linear QSAR models, therefore an objective comparison should determine the most suitable approach.

Data sets are the most important aspect to be considered when constructing a predictive model, because the final accuracy is directly related to the consistency of information that is processed. From the experimental images provided by the medical research team, various features can be extracted using image processing tools. There are multiple methods that can be used, from traditional to deep-learning techniques, all with excellent results concerning cell segmentation. After information gathering, the molecular descriptors are computed in order to determine the set of significant features that best predict the biological activity values. The selection of these features represents also the optimization phase. The novelty of the work lies in application of suitable image processing techniques that match the necessities best, before proceeding to model development. The provided images are pre-processed for quality enhancement, adjusted morphologically to achieve smoothness and more regular shapes, then using watershed segmentation on imposed regional minimums cells are being labeled such that features can be extracted efficiently. Further, the ANFIS model is developed and optimized using Antlion optimization algorithm.

The following section of the current study will cover the methods and algorithms chosen for implementation of both image segmentation and model development, some preliminary results regarding the processing of fluorescent images and finally conclusions for further advancement.

2. METHODS

2.1 Image Processing

The dataset used is provided by the research team from "Iuliu Haţieganu" University of Medicine and Pharmacy. After several experiments conducted, some of their results consist of fluorescent image sets that depict the effect of treatment applied to cell cultures, the reaction of cells for different doses (Fig. 1).

A major impact of image processing in this study is to extract and measure numerical information about vaccine efficiency, further used as parameters for the model design. These parameters could include cell luminosity, surface, shape, or counting, depending on each one's significance for the desired model.

The fundamental aspect that needs to be reached is the cell segmentation, i.e. splitting of microscopic image into regions representing each cell and differentiate from the background. The main objective is to acquire data about individual cells or cells grouped as a whole but separated from the background.

The approached process is a combination of traditional adjustment methods and watershed transform. Given the present quantification problem, this technique is quick and straightforward, and at the same time offer great results. In contrast with the deep learning methods that can be used for segmentation with minimum human intervention, the classic ones have the downside of manually selecting each step of the

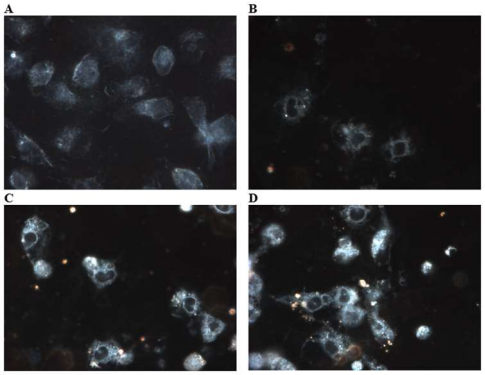

Figure 1. Cell exposure to various drug quantities

algorithm based on visual outcomes. This can lead to potential human errors. The segmentation can be performed through various procedures, among which edge detection commonly used in cell determination. Although, for this particular problem, shapes are highly irregular, and cells strongly condensed, hence contours cannot be recognized properly with either edge detection or machine learning techniques if the dataset is not large enough. On the other side, watershed segmentation is addressed successfully in diverse object separation problems (Chourasiya & Rani, 2014).

A first step in the algorithm is the selection of the green band from the 3-channel RGB picture, because the original acquired image captures green fluorescence on a black background. High performance laboratory equipment provides qualitative images with full brightness such that an illumination uniformization is not strictly necessary. However, contrast adjustment is a helpful improvement essentially for developer analysis, since the conversion from color image lead to a dimmed image. For the proposed implementation, an adaptive method is used to increase contrast equalization, that is *adapthisteq*. The major leverage it has, in contrast with global enhancement methods is that this function processes small portions of the image based on histogram assessment, preventing noise amplification (Gonzalez, et al., 2009). Prior to adjustment of contrast, a median filter must be applied with the aim of reducing the noise potentially infiltrated in the original image. A median filter has more effective results whilst preserving edges, by computing a statistical median of neighbor pixels in comparison with mean filtering.

The image is further binarized using a global threshold determined by Otsu method and then morphological opening and closing are applied to enhance smoothness of all edges detected (Sobhy, et al., 2016). The smaller stains which are not conclusive in regard to the cell segmentation can be considered noise and removed such that the number of cells established from the segmented image to not be influenced. On the same idea *bwareaopen* can be used also for cleaning the small disturbances inside the cell area, through complementing, because cells represent the foreground.

The watershed segmentation is based on the concept that each object we want to identify should be a catchment basin, meaning we need to transform the image using gradient and from there a selection of local minima is chosen. Each minimum will correspond to a pool that represents the cell region, computed using distance transform (Gonzalez, et al., 2009), connected as a graph. The watershed can finally be applied, but with the observation that typically over-segmentation may appear due to the fact that each and every local minimum is settled to be a catchment region, hence some parts of cell are segmented inconsequentially (Ji, et al., 2015). By combining *imextendedmin* and *imimposemin* functions the issue is solved through finding deep regional minima and imposing these as cell nuclei. Finally, once more, spots smaller than a selected threshold are removed because cannot be considered whole cells.

2.2 QSAR Vaccine Model

Data sets of chemical descriptors describe the composition, structure and behavior of evaluated compounds proposed by the vaccine developers. A variety of molecular descriptors, describing constitution, topology, thermodynamic behavior, geometry, etc. are determined from experimental data conducted in the laboratory (Mocan, et al., 2015).

As stated previously, due to molecular relationships that are highly nonlinear, the focus is directed towards nonlinear model approaches like artificial neural networks. Precision of ANNs depends on parameters therefore can end up trapped in local optima. An improved alternative is the combination of ANN with fuzzy logic rule set, resulting the Adaptive Neuro-Fuzzy Inference Systems (ANFIS). The fuzzy logic system introduces a subjective human reasoning to the machine learning algorithm of ANNs. The first fuzzy method belongs to Takagi and Surgeon (Takagi & Sugeno, 1985) who defined a set of fuzzy rules for creating a nonlinear association of inputs with the outputs, in the form *IF premise THEN consequent*. A downside of ANFIS method is the impact of employing numerous descriptors, which in medicine and biological fields are commonly used. More specific, the complexity of the net would increase in such a way that it could lead to overfitting issues determined by the trained parameters, whilst the overall accuracy would be diminished (Elaziz, et al., 2018).

Considering the highest accuracy is desired for any QSAR model, one essential step is to make a rigorous selection of the most significant describing features that will determine the model and implicitly the biological activity prediction. A new concept inspired from the natural hunting mechanism between antlions and ants, the antlion optimizer (ALO) algorithm which was advanced firstly in (Mirjalili, 2015), can be used for selection of optimal descriptors. The positions of ants and antlions are initialized randomly, the ants then simulate a random walk until a trap occurs. Another observed correlation appears between the antlion fitness and the hole dimensions because the chance of an ant being caught increase with the pit radius. The fitting is computed using (1), taking into account the mean squared error and the number of features chosen versus the entire set. When an ant fitness is greater than the antlion's, it means that the ant was captured by the antlion hence the last position of the ant becomes the new position for the trap in order to increase the next catch probability.

$$f_i = \alpha \times \sqrt{\frac{1}{N_S}\sum_{i=1}^{N_S}(\hat{Y}_i - Y_i)^2} + (1-\alpha) \times \left(\frac{|x|}{D}\right) \qquad (1)$$

The main goal of this optimization algorithm is to minimize the final error and to adjust the number of advantageous features. The resulting set of features is further used as input descriptors for the ANFIS model to predict biochemical activity. An improved version is the "*Tournament antlion optimization algorithm*" (TALO) (Kılıç & Yüzgeç, 2019) because the size of the ant's random walk is decreased five times to diminish runtime. It's second advance concerns the equations that model the antlion's pit and the slip rate into the trap. The newly upgraded method assures that the ant's positions does not exceed the search region by introducing limits to the ant model equation.

In general, the ANFIS architecture is designed as a neural network of five layers with the neurons from each layer having the same function, all part of the fuzzy inference process depicted in Fig. 2: computing the fuzzy values, firing strength of fuzzy rules, normalize firing strength, combine premises

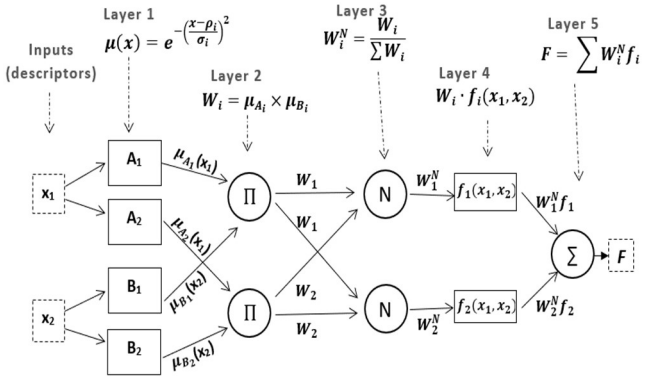

Figure 2. ANFIS Architecture

with consequents, predict and final output. Optimization is achieved through the weight parameters between layers.

The fuzzification process, which is represented as first layer, consist in determination of fuzzy values from the inputs received using membership functions (MF). These MFs are indicators of values membership to a certain category, such that the data inputs are divided into distinct categories in a low to high range with a mean value and the deviation which implies the similarity degree between these values. Specifically, the membership functions assign values for the inputs by using bell-shaped Gaussian function, according to the category they belong to. Each input will be described by a membership value to each category such that finally, in this layer, the number of nodes is n × m, i.e. n is the number of members (inputs) and m is the number of categories.

The firing strength from layer two is in essence a weight computed by employing the previously fuzzified values, so

each weight depicts the strength of the respective rule from layer one. The weights are normalized in layer three, in this way each weight being compared to the other ones, such that the larger the strength, the better is the rule.

In the fourth step, the computed weights are combined with the

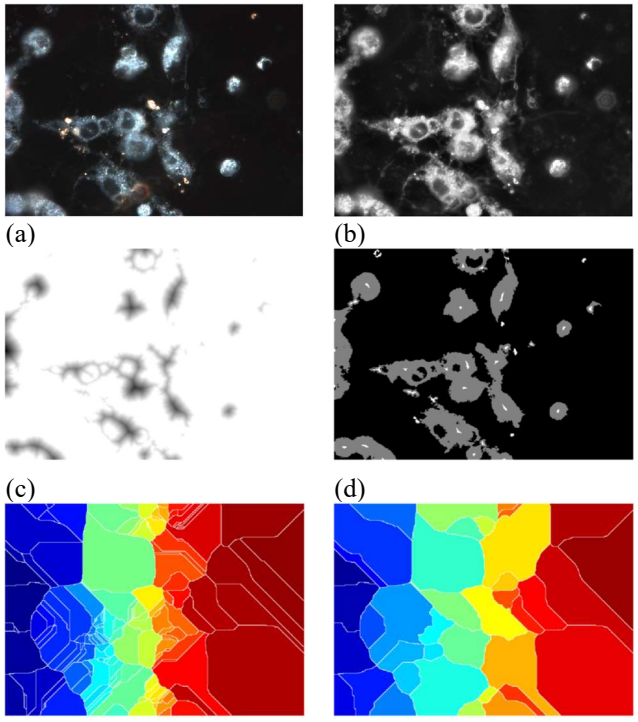

Figure 3. (a) original image (b) contrast adjustment (c) catchment basins (d) nucleus mask over binary image (e) watershed transform of c (f) watershed transform of c with d imposed

input data again to result the consequent function and the corresponding output values that are summed up in the fifth and last layer to determine the predicted activity. The learning process comprises two techniques, forward pass and backpropagation. The inputs are feed into the net and on forward pass the consequent parameters are developed using least squares estimation (Jang, 1993) and after the error is determined, through backpropagation, from output back to input by gradient descent, the parameters for premise are settled.

Consequently, the model proposed would be a combination of ALO-ANFIS that in the first stage selects the descriptors to be used as inputs for the ANFIS prediction of inhibitors activity, in second stage.

3. RESULTS AND DISCUSSION

The procedures are implemented in Matlab by using the Image processing toolbox for segmentation and run on a 64-bit Windows environment.

3.1 Image Segmentation

The point of interest for this research is the difference in images by comparing stages with applied vaccine doses to those control stages, where cell culture is raw. An analysis of reflection points from cells and on the membrane should conduct to first conclusions that can be extracted, to confirm the effective presence of the bio-nanocomposite inside molecular area.

Each image runs through the series of pre-processing steps mentioned previously to ensure a high quality and proper segmentation (Fig. 3b), including noise reduction, contrast enhancement and morphological opening-closing. Images presented in Fig. 3 are acquired through processing steps proposed, using Matlab.

After that, the bright areas are turned into catchment basins (Fig. 3c). A mask comprising cell nuclei (Fig. 3d) is created by computing regional minima to be further imposed on the binary image to avoid over-segmentation, depicted in Fig. 3e, due to the consideration of all local minima found as potential nucleus. In Fig. 3f, the watershed segmentation is applied on the same catchment basins image, but this time regional local minimum is imposed to determine more accurate splitting.

Figure 1 illustrates the impact of different doses on cell culture. Stage A represents the cells in their initial condition, in stage B a small quantity of vaccine is administered and further to C, D the concentration is raised gradually. As it is observed, the higher the concentration, the greater the number of bright spots, consequently the effect of the treatment is more accentuated. Following the same direction, the number of clear white spots or also the surface covered is directly connected to efficiency on respective case.

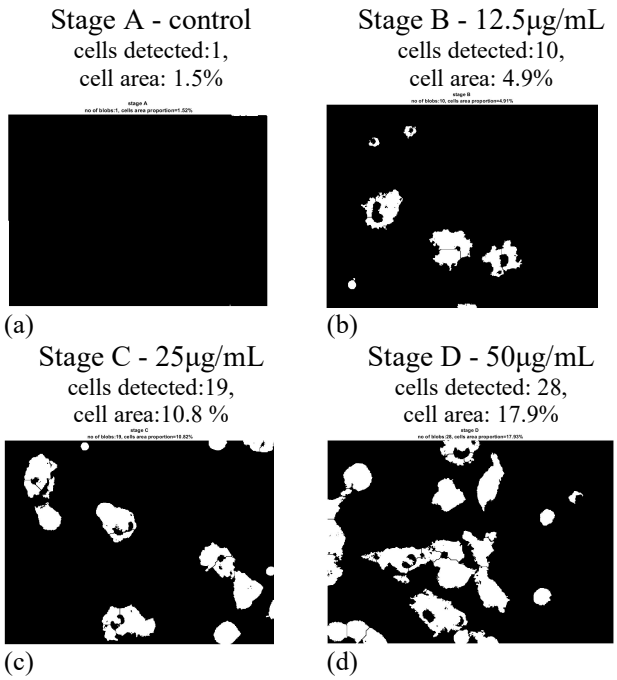

Figure 4. Segmented images

A processed set is presented in Fig. 4, as well as some extracted features for the cell group captured. Having the binary image labeled after segmentation, an advantageous tool is *regionprops* for obtaining significant cell related numerical data like area, perimeter, roundness degree or extrema points. In a structure element each cell counted will correspond to data of interest for assessing viability and proliferation of cells,

their whole surface is computed as a rate over entire image area to have a plain contrast from one stage to the other. The molecules tend to form clusters, especially when are exposed to higher concentration doses.

The histograms are created from original images (Fig. 5) and indicate strong differences between phases, especially D versus A, where visibly the number of high intensity pixels is substantially greater, values beyond 200 in the range. The distribution is right-skewed because the black pixels from dark-field background are predominant in all images.

For a final inference, every result in a group of four dose-dependent images is compared to phase A, that is without valuable information, it contains just noise captured from the microscope system. The interest is directed towards modifications from zero point, i.e. stage A.

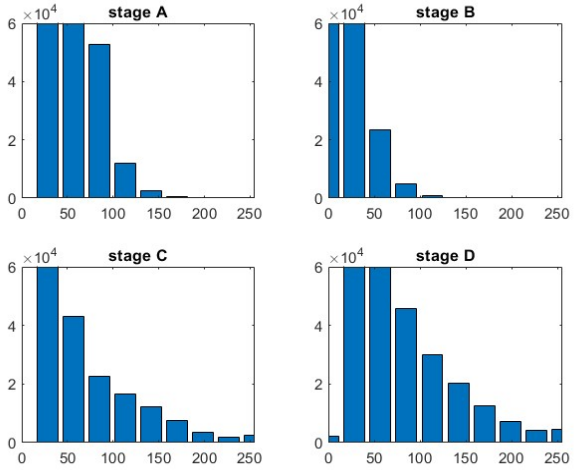

Figure 5. Histograms for a group of images

3.2 Validation of the methods

One of the main resolutions is the quantification of single-cell characteristics, therefore it is compulsory to validate the accuracy of the proposed method for cellular identification. To establish this, manually counted cells were used as reference point for comparing the segmentation results with the measures. Hence for assessing the results, firstly the binary masks determined by segmentation method are determined. Then, each result is evaluated by human perception. To achieve this, 28 different test captures were used to test over 1300 cells. For each image *i*, error values were computed to measure the accuracy rate of the segmentation. The error rate can be determined using (2), where $B\left(b_j^{cell}\right)$ is the binary indicator which takes 1 and 0 values according to the masks overlay. Specifically, if b_j^{cell} (mask pixels of cell *j* after application of the proposed method) belong to $b_j^{cell,Original}$ (mask pixels of cell *j* in the human labeled image *i*) then $B\left(b_j^{cell}\right) = 1$, else $B\left(b_j^{cell}\right) = 0$. N is the total number of cells in a frame *i*.

$$\varepsilon_i = \frac{\sum_{j=1}^{N} B(b_j^{cell})}{N} \times 100\% \qquad (2)$$

The resulted error rates indicate an error mean of 10.6% with a standard deviation of 2.4% constituting a great result for cell area definition.

Each image corresponds to a category depending on the concentration of treatment applied (0, 12.5, 25, 50 μg/mL) and the determined values are stored in a table with four distinguished rows, such that a visual representation of percentual number of cells determined by the segmentation method and by human labeling is depicted in Fig. 6.

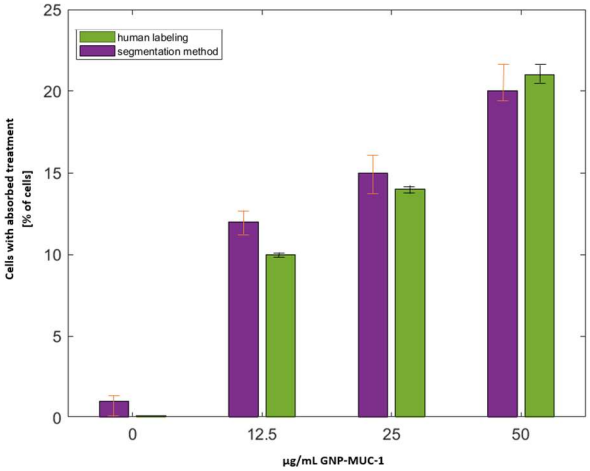

Figure 6. Determined cells in each concentration category manually versus through proposed segmentation method

Then the mean absolute error (MAE = 0.287) also confirms that the variation between the reference number of cells and determined cells is not significant. Therefore, the proposed algorithm can be successfully applied on cell populations in the ongoing study for single-cells determinations.

The experiments are still in progress such that the model is yet in preliminary process. However, for validation of the developed models, both internal validation (or cross-validation) and external validation will be applied. The obtained model will be used both in prediction and optimization steps. The split of data set into training and test sets is a vital step because it must cover the entire information that was gathered, i.e. the entire data set. The training set is used for model development whilst test set will evaluate to what extent the model predicting is effective and reliable. There are many ways in which we can consider the splitting, each one having different performance levels. The goodness-of-fit assess how well the model predicts the training set. For internal validation, there is cross-validation, used also in (Elaziz, et al., 2018), where they used a k-fold cross-validation, more specific, one class from a set of k equal in size classes of data, was used, at a time, for testing while the other k-1 were used for training, and this procedure was applied k times. Besides, as external validation, the whole set is divided in two specific parts, for most of the cases using a proportion like 80-20% (or 70-30%) training and testing sets, respectively. In this case the objective is how well the model predicts the test set.

By applying Leave-One-Out Cross-Validation firstly on training set, one can determine the models with potential

predictive results, the ones with a cross-validation coefficient q^2 over the imposed threshold of 0.6 encountered in literature. These models were then verified on internal test sets such that the squared correlation coefficient R^2, signifying the difference between measured and predicted activity, to reach values higher than 0.6. If both conditions are satisfied, the corresponding models could be further validated on external test datasets. For the assessment of the resulted model the best practice would be to employ several types of measures such as statistical, prediction accuracy and domain applicability. The statistical methods include correlation coefficient R^2, mean squared error and root mean squared error. Statistical parameters evaluate the fitness of the model; however, it does not ensure the prediction level to be high. Moreover, multiple inequalities involving R^2, R_0^2 (correlation coefficient through the origin) were utilized to evaluate if the prediction has adequate accuracy on both training and testing sets. The difference between the two coefficients is that R_0^2 verifies the actual accuracy of the prediction, how well data fits to the model, while R^2 analyze the correlation between real and computed activities.

Each method has its advantages and drawbacks such that many things must be considered to avoid overoptimistic results: diversity of data, statistical methods, sizes, study purpose. By comparing the results from all three different approaches we can observe the problems in model structure.

4. CONCLUSIONS

The purpose of the model is to assess safety and efficacy of the vaccine nano-compounds. The image processing phase is relevant for feature extraction regarding pixel values but also independent cell information to determine reaction level of different molecules exposed to the therapeutic vaccine. The obtained preliminary results are conclusive for further analysis; hence the extracted features will be engaged first in the feature optimization process and then the selection will determine a desired model through the neural network constructed.

The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Ethical Committee of the Iuliu Hatieganu University of Medicine and Pharmacy (no.16/21 January 2021).

ACKNOWLEDGEMENT

This research was funded by a grant of the Romanian Ministry of Research and Innovation, CCCDI—UEFISCDI, project number PN-III-P2-2.1-PED-2019-0844, contract no.323PED/2020.

REFERENCES

Chourasiya, S. & Rani, G. U., (2014). Automatic Red Blood Cell Counting using Watershed Segmentation. *International Journal of Computer Science and Information Technologies,* 5(4), pp. 4834-4838.

Elaziz, M. A., Moemen, Y. S., Hassanien, A. E. & Xiong, S., (2018). Quantitative Structure-Activity Relationship Model for HCVNS5B inhibitors based on an Antlion Optimizer-Adaptive Neuro-Fuzzy Inference System. *Scientific Reports,* 8(1).

Gonzalez, R. C., Woods, R. E. & Eddins, S. L., (2009). *Digital Image Processing using MATLAB.* s.l.:Gatesmark Publishing.

Jang, J.-S. R., (1993). ANFIS: adaptive-network-based fuzzy inference system. *IEEE Transactions on Systems, Man, and Cybernetics,* 23(3), pp. 665-685.

Ji, X. et al., (2015). Cell Image Segmentation Based on an Improved Watershed Algorithm. *8th International Congress on Image and Signal Processing,* pp. 433-437.

Kılıç, H. & Yüzgeç, U., (2019). Tournament selection based antlion optimization algorithm for solving quadratic assignment problem. *Engineering Science and Technology, an International Journal,* 22(2), pp. 673-691.

Mirjalili, S., (2015). The Ant Lion Optimizer. *Advances in Engineering Software,* Volume 83, pp. 80-98.

Mocan, T. et al., (2015). In Vitro Administration of Gold Nanoparticles Functionalized with MUC-1 Protein Fragment Generates Anticancer Vaccine Response via Macrophage Activation and Polarization Mechanism. *Journal of Cancer,* Volume 6, pp. 583-592.

Nancy Moyer, M., (2019). *Adenocarcinoma Symptoms: Learn Symptoms of the Most Common Cancers,* s.l.: Healthline.

Peter, S. C. et al., (2019). Quantitative structure-activity relationship (QSAR): modeling approaches. *Encyclopedia of Bioinformatics and Computational Biology,* pp. 661-676.

Pradeep, P., Friedman, K. P. & Judson, R., (2020). Structure-based QSAR models to predict repeat dose toxicity points. *Computational Toxicology,* Volume 16.

Sobhy, N. M., Salem, N. M. & Dosoky, M. E., (2016). A Comparative Study of White Blood cells Segmentation using Otsu Threshold and Watershed Transformation. *Journal of Biomedical engineering and medical imaging,* 3(3).

Sung, H. et al., (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *A Cancer Journal for Clinicians.*

Takagi, T. & Sugeno, M., (1985). Fuzzy identification of systems and its applications to modeling and control. *IEEE Transactions on Systems, Man, and Cybernetics.*

Tang, H. et al., (2009). *Novel Inhibitors of Human Histone Deacetylase (HDAC) Identified by QSAR Modeling of Known Inhibitors, Virtual Screening, and Experimental Validation,* s.l.: American Chemical Society.