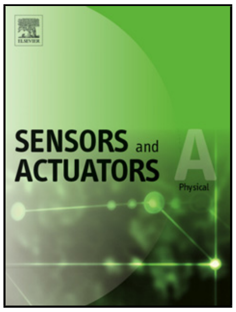


Low cost color assessment of turbid liquids using supervised learning data analysis – Proof of concept

Daniel P. Duarte (Conceptualization) (Methodology) (Software) (Formal analysis) (Writing - original draft), Rogério N. Nogueira (Writing - review and editing), Lucia B. Bilro (Validation) (Writing - review and editing) (Supervision) (Funding acquisition)

PII: S0924-4247(19)31755-8

DOI: https://doi.org/10.1016/j.sna.2020.111936

Reference: SNA 111936

To appear in: *Sensors and Actuators: A. Physical*

Received Date: 23 September 2019

Revised Date: 21 January 2020

Accepted Date: 2 March 2020

Please cite this article as: Duarte DP, Nogueira RN, Bilro LB, Low cost color assessment of turbid liquids using supervised learning data analysis – Proof of concept, *Sensors and Actuators: A. Physical* (2020), doi: https://doi.org/10.1016/j.sna.2020.111936

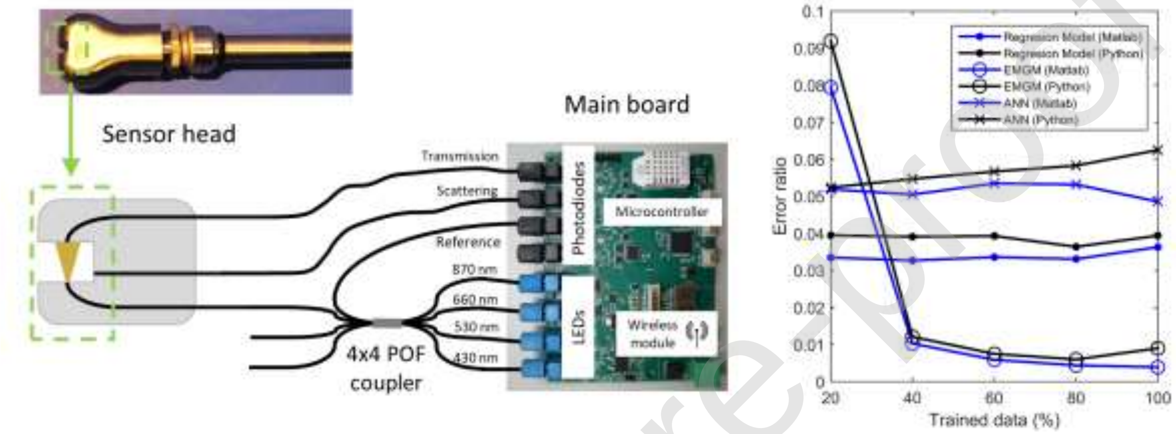
Low cost color assessment of turbid liquids using supervised learning data analysis – Proof of concept

Daniel P. Duarte[*,a,b], Rogério N. Nogueira[a], Lucia B. Bilro[a]

[a] Instituto de Telecomunicações – Polo de Aveiro, Aveiro 3810-193, Portugal
[b] Department of Physics, Aveiro University, Aveiro 3810-193, Portugal

[*]Corresponding author: dduarte@ua.pt

Graphical_abstract



Highlights

- Low cost in-line color sensor for turbid liquids developed.
- Color and turbidity successful discriminated using three analytical approaches.
- Regression models is best suited for standard or occasional measurements.
- Expectation Maximization Gaussian mixture performs better for well-known controlled range of colors and turbidities.
- Artificial neural networks have easy implementation and is suited for real-time Internet of Things platforms.

Abstract

This work reports the development of a low cost in-line color sensor for turbid liquids based on the transmission and scattering phenomena of light from RGB and IR LED sources, gathering multidimensional data. Three different methodologies to discriminate color from the turbidity influence are presented as a proof of concept approach. They are based in regression models, expectation maximization Gaussian mixtures and artificial neural networks applied to labeled measurements. Each methodology presents advantages and disadvantages which will depend on the intended implementation. Regression models revealed to be best suited for standard or occasional measurements, the EM Gaussian mixture will perform better for well-known controlled range of colors and turbidities and the neural networks have easy implementation and potential suited for real-time IoT platforms.

Key words: Color, Turbidity, Sensor, Artificial Neural Network, Expectation Maximization Gaussian Mixture, Clustering

1. Introduction

Fluids found in nature and the ones used or produced in industry generally present, at some degree, suspended particles of solid matter with different sizes. If the liquid is in regular agitation, the particles will never settle down, making the liquid appear turbid. When color determination is necessary for quality control, laboratory analysis consisting of centrifugation and spectroscopy measurements are usually performed [1–4]. This type of analysis requires the extraction of a sample, its transportation to a laboratory where costly equipped such as a centrifuge and the spectrometer are located and, ultimately, the waste of the sample. This is, overall, a very time-consuming process.

Low cost methods to measure turbidity in water samples were developed by Omar and MatJafri [5] using plastic optic fibers (POF) to guide light from 470 and 633 nm LED to a measurement cell where a water sample is placed. Light from both 0º and 90º is detected with photodetectors using other fibers as waveguides. The light intensity varies with the turbidity level. Bilro et al. uses the same concept and takes the advantage of POF miniaturization and transportability to develop an in-line solution, but using a 660 nm LED [6]. These solutions are color sensitive, but do not perform color measurements and thus unfitted for colored liquids. Another in-line solutions were developed and tested by Garcia et al. [7] and Crespo et al. [8] having the same principle as Omar and Bilro proposals, but in this case an infra-red (IR) LED was used with the propose to be color insensitive. Although turbidity in colored liquids, that do not absorb in IR, is calculated with this approach using an IR LED, the determination of color or spectral bands of interest is not possible. A more recent proposal using multimode fibers was presented in 2019 by Yeoh et al [9]. It attached side by side two fibers with their beveled tips mounted vertically and due to evanescent light transference between the two fibers in the turbid medium, light intensity variations were observed. However low accuracy values were obtained. Other low cost approach using smartphone as a detection device was proposed by Hussain et al. [10] where a system with a sample holder and IR LED is coupled with the smartphone. This system uses the smartphone battery for powering the IR LED that illuminates the sample holder for a nephelometric 90° measurement that is done with the smartphone IR sensor (ambient light and proximity sensor).

For clear colored liquids, Jiménez-Márquez et al. [11,12] proposed a low-cost solution for a color sensor using transmitted light of LEDs with wavelengths of interest. The light is targeted to a measurement cell and the transmittance measured through a photodiode. Novo et al. [13] used the same approach but using POF as light guidance to the measurement cell. A smartphone based color determination sensor was also developed by Sumriddetchkajorn et al. [14] for chlorine concentration assessment where a self-referencing analysis is done for converting the color level of water to its corresponding chlorine concentration. A portable closed chamber is used as the support structure for the water sample and smartphone, with the concentration calculated from the ratio between the color intensity of the water sample and the empty sample holder. This process requires prior reagent mixing for chlorine activation. All the solutions here presented were not in-line prepared neither turbidity insensitive.

In this manuscript is presented an in-line solution that uses not only the IR LED, but also colored red, blue green (RGB) LEDs. The light will be guided using POF as the waveguide to a measurement cell with the detection occurring at both 0º and 90º for each LED. These multivariate data will be used for the determination of color from turbid liquids. Because both turbidity and color interact with the sensor’s optic system in similar way, a model and statistical correlation analysis must be done to discriminate each parameters effect. Machine learning algorithms will be used, as a proof of concept approach, and a comparison stating the advantages and disadvantages of each algorithm is performed. Simple calibration regression models, expectation maximization Gaussian mixture (EMGM) [15–17] and artificial neural networks (ANN) [18,19] are used to correlate prior labeled data, this is, using supervised learning. The comparison will be performed considering features like easy of application, time of processing and error estimation by using random sampling cross validation.

2. Sensor Structure

The sensor here presented is based on the 0º transmission and the 90º scattering light measurement. Previous work on this concept can be consulted in [6,20,21]. The sensor has four LEDs sources from Industrial Fiber Optics, Inc. with central wavelength of 430 nm (IF E92), 522 nm (IF E93), 660 nm (IF E97) and 870 nm (IF E91D) with bandwidth of 65, 40, 40 and 50 nm respectively. These RGB wavelengths were used based on the Glories method color system [22]. The LEDs are controlled by a main board with microcontroller for analog to digital converter and a wireless connection module. Each LED is connected to the same side input of a 4x4 POF coupler. Only two output fibers of the coupler are used, one as a reference/compensation for the LEDs electrical drift and the other will be the waveguide of light to the measurement partition. This partition is part of the measurement head sensor that can be submersed for in-line measurements. Two other fibers are connected to the partition in the same plane of the emission fiber but at 0º and 90º. These are the ones responsible to receive the scattered and transmitted light respectively. The 0º fiber is at a distance of 5 mm from the emission fiber and the 90º is right next to it. These fibers are then connected individually to a photodiode (IF D91 from I. Fiberoptics) in the main board. In total, counting with the reference fiber, three photodiodes are used (Figure 1). To obtain a measurement, the main board turns on and off individually each LED in sequence and registers the output signal as voltage (*V*) obtained by the photodiodes which is proportional to the intensity of light. In total 4 scattering values and 4 transmitted values (8-dimensional data) will be obtained.

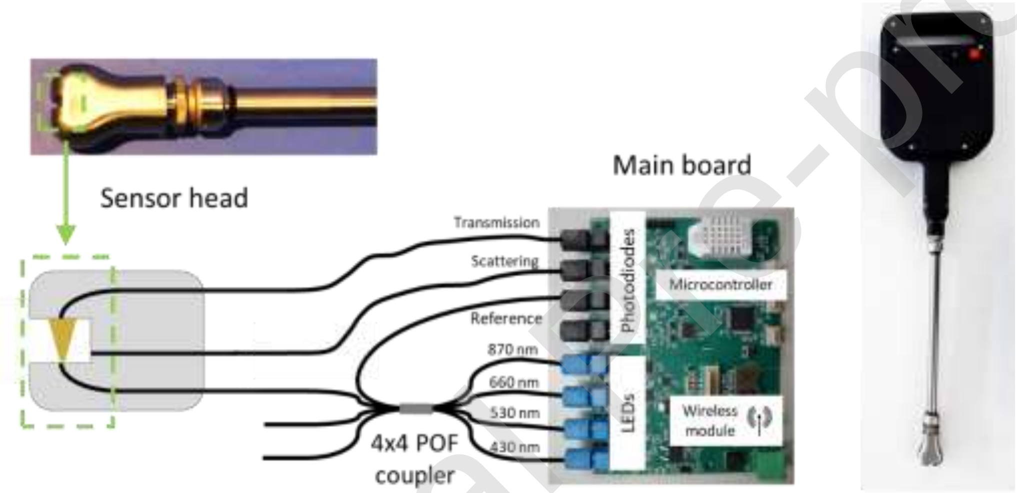


Fig. 1. Developed in-line color sensor schematic with turbidity insensitivity. All the measured data is transmitted by the wireless module. Full image of the assembled sensor in the right.

3. Measurement samples

Samples of 500 ml with different colors and turbidities were prepared using red, yellow and green food dyes and corn starch at different concentrations. The volume of each dye added to water was 0.5, 1, 1.5 and 2 ml. Spectral absorbance analysis for each dye concentration was performed in the most relevant window and is presented in Figure 2.

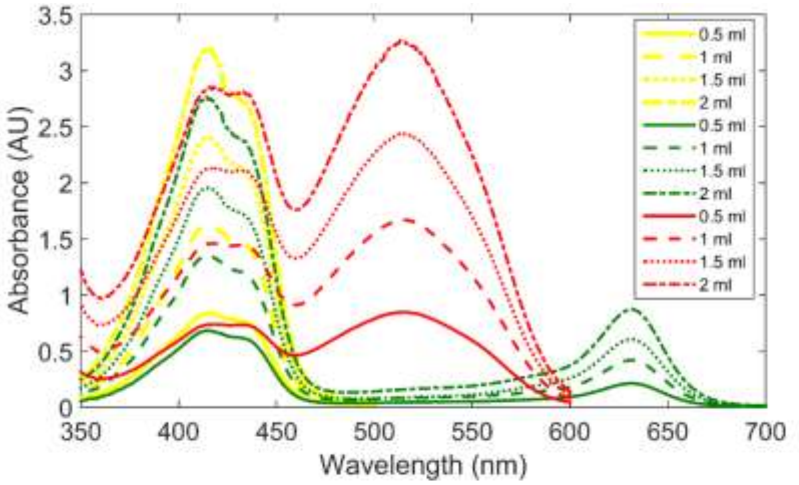


Fig. 2. Absorbance spectral analysis for each colored solution using food dyes at different concentrations. The yellow, green and red lines are related to the yellow, green and red dyes used respectively.

For each color, mass combinations of 0, 0.2, 0.45, 1, 2, and 2.8 g of corn starch was added and mixed. The dyes do not absorb in the IR region of the related LED. The turbidity value of each sample was measured by the commercial turbidimeter from Libelium S.L. [23] (calibrated with formazine standards and has a 5% accuracy) and a range up to near 4000 NTU was chosen to test the sensor for turbidities that could be of interest in same industrial applications such as the beverage manufacturing. A total of 78 different samples were obtained. Table 1 resumes the prepared samples.

Table 1
Resume table of the prepared samples used to train and classify.

	Dye volume (ml)	Turbidity (NTU)					
No color	0	0	164	449	999	2020	3571
Yellow dye	0.5	0	187	461	1042	2095	3542
	1	0	175	464	1016	2158	3556
	1.5	0	182	468	1033	1989	3456
	2	0	198	462	1052	2080	3575
Green dye	0.5	0	187	415	1021	2065	3721
	1	0	174	431	999	1973	3518
	1.5	0	183	424	1044	2053	3535
	2	0	196	438	1010	2089	3530
Red dye	0.5	0	204	474	996	2176	3688
	1	0	222	450	1023	2143	3529
	1.5	0	203	465	1023	2049	3530
	2	0	205	420	980	2006	3534

With all the samples prepared, a cycle of 3 different measurements were taken for each sample lasting 3 minutes each. A measurement of 3 minutes has a total of 8 individual points that has a Gaussian distribution intrinsic to the sensor measurement error and the Brownian motion of the particles. Figure 3 shows an example of transmittance and scattering mean values of the measurement of each red dye done using the 522 nm LED. All the values obtained are normalized with the measured voltage obtained from clean water (V_0).

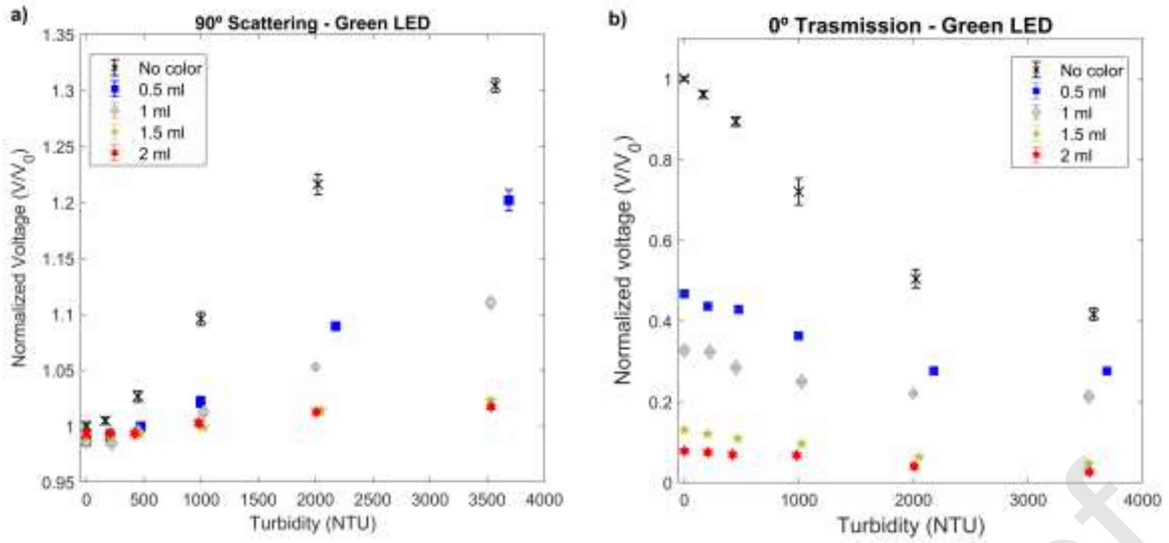


Fig. 3. Measurement for each red dye obtained for the a) 90º scattering and b) 0º transmission green LED. For higher concentrations of dye, it is possible to observe the influence of its light absorbance nature.

With the increase of turbidity and/or color, the transmitted light will diminish with the growing number of scattering particles and/or absorption centers. This decreasing of light intensity (*I*) is ruled by the Beer-Lambert law:

$$I = I_0 \times 10^{-\alpha l c} \qquad (1)$$

where I_0 is the initial light intensity, α the molar attenuation coefficient, l the optic path and c the concentration of the attenuation substance. By using the RGB LEDs, it is possible to characterize the color from the transmittance values. Note that each attenuation agent will lower each other sensitivity. Therefore, if the color concentration is increased, the sensitivity for turbidity is lowered and in the same way, if the turbidity of the solution is increased, the sensitivity for color detection is also lowered.

The scattered light that reaches the 90º fiber increases with the concentration of turbidity and it can be described by the Mie solution to Maxwell's equations [10,24] which complex formalism can be found in [25]. A 2nd order polynomial equation can be used as regression model. Color in the 90º scattered light will have the same behavior as transmission because of the same Beer-Lambert absorption phenomena. With all of this data classified, different approach of data analysis can be performed to create a good color classifier for turbid liquids.

4. Data analysis

Using the labeled data obtained with the sensor from the samples, here it will be presented three ways to analyze the data and to discriminate the contributions of color and turbidity. This study tries to present a different view as a proof of concept to how to analyze the multivariate data received from a sensor with the aim to obtain a better measurement. The first method will be based on simple regressions of mean values and will take advantage of the IR LED transmission insensitivity of color to calculate first the turbidity and then the color. The second method will take the advantage of the measurements' Gaussian distribution to create clusters that will be the elements in cluster-continuum regression models with the variation of turbidity and color. The last method will use artificial neural networks (ANN) that will create a non-linear model that could discriminate the color of the liquid independently of the turbidity value.

4.1 Regression models

This method is the simpler and most direct one basing its prevision from the regressions performed to the non-color turbid data. Therefore, it doesn't need the labeled data from the dye samples. The regressions discussed in 3, this is, the 2nd order polynomial for the scattering data and the Beer-Lambert for the transmitted data, are applied to the non-color solutions to be trained (Figure 4). Having the trained regressions, it is possible to assess the color of a new measurement by finding the turbidity estimative of the sample through the IR LED information, which is color unresponsive (Figure 5). This can be done by using the scattered regression, the transmission regression or a combination of both. Because of the lower error presented for higher turbidities from the transmission regression, this is the one chosen to infer the turbidity obtained. By knowing the turbidity value, the color determination is performed by using the transmitted regression from the other LEDs. First, we find for each LED the output signal value that this turbidity value was supposed to have if color absorption were not existent (V_{nc}). This is the new initial light intensity signal output value that will be used as the base value to calculate color. Then we compare it to the measured value (V_m) from the unknown sample and calculate the transmittance through T_r=V_m/V_{nc}. It is obtained in this away three RGB values of transmittance that characterize the color of that unknown solution.

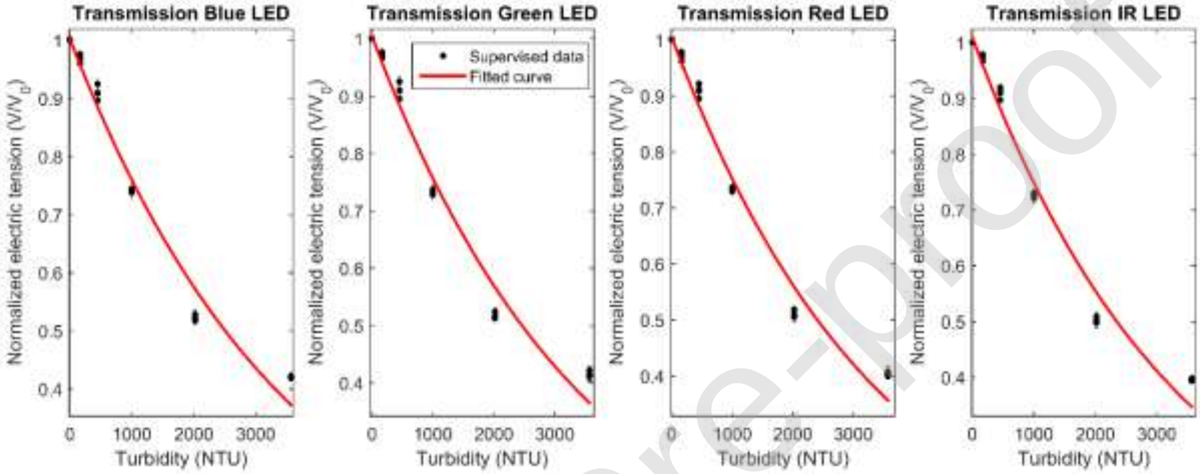


Fig. 4. Transmission Beer-Lambert law based regression for normalized supervised data obtained from all the LEDs of the sensor.

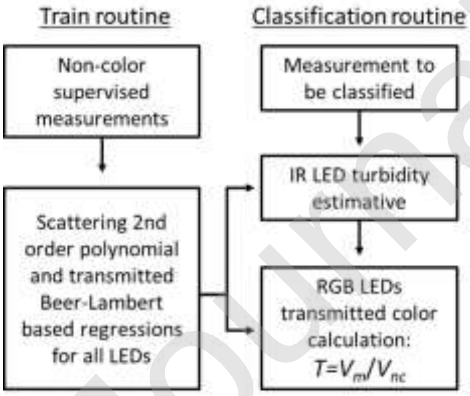

Train routine
Classification routine
Non-color supervised measurements
Scattering 2nd order polynomial and transmitted Beer-Lambert based regressions for all LEDs
Measurement to be classified
IR LED turbidity estimative
RGB LEDs transmitted color calculation:
T=Vm/Vnc

Fig. 5. Train and classification routines using regression models to the supervised data, taking advantage of the IR LED color insensitivity for RGB color determination.

4.2 Expectation maximization supervised Gaussian mixture

The second method will take advantage of the intrinsic Gaussian distribution that a measurement taken over time with several points will have around a mean value. For this, an expectation maximization supervised Gaussian mixture (EMGM) algorithm is used [15,16]. The EMGM algorithm will pick all the supervised data to be trained and will find the maximum likelihood solution to create a prechosen number of clusters, considering its Gaussian distribution (Figure 6).

For each cluster, a multidimensional mean value, covariance matrix and a mixture coefficient will be calculated. In this method, a cluster will be calculated for each unique set of turbidity and color present in the randomly selected training set, which ensures that a single cluster representation is performed, this is, the number of clusters is optimized with dependence with the training set. This also prevents that an outlier will be considered to its original cluster measurement since it will be integrated to a closer cluster, diminishing the overall error. Having each measurement associated to a cluster, the same regression models used in the first method are applied to the mean values of the clusters but, in this case, it is applied to all of the colored solutions as well. This process can be seen as a creation of a cluster-continuum for each colored solution that varies with turbidity and that has not a single mean value but instead a "continuous" mean value taken from the regression models applied. In this cluster-continuum, a global variance will be considered as the mean variance values of the prior clusters that constitute the cluster-continuum and the mixture coefficient will be the sum of their individual mixture coefficients. With this approach, independently if it was already trained or not, any turbidity value can be associated to a sub-cluster from the cluster-continuum. Finalized the training routine with the determination of each cluster-continuum for each color. The classification routine initializes by estimating the turbidity value of a measurement using the IR regressions like in the 1st method presented in 4.1. In alternative, an approximate value of turbidity, using only the transmission and scattering dimensional information of the RGB LEDs, can be calculated based in a recursive cycle that minimizes and compares the error of the expected turbidity for each individual dimension, which makes this approach IR independent but increases the overall error. This turbidity estimation will be very important because it will be the value that will be used to calculate, for each color cluster-continuum, the comparison multidimensional point that will be used later. In other words, the estimated turbidity value will be used to calculate the expected value for each dimension using the regressions obtained by the training routine. Doing this for each color, in the end we will have the measurement point that we want to determinate and, for that expected turbidity, the expected sub-clusters for each color that encompasses the multidimensional mean point, the covariance matrix and the mixture coefficient. It is now possible to determinate the weight or how close is the measurement point to the sub clusters of each color. The more weigh that a sub-cluster has with the measurement point, the higher probability of that point to be of that sub-cluster color. If the calculated weight is mostly distributed between two sub-clusters, then the probable color will be between those two sub-cluster colors. The color is therefore calculated based on a percentage of each sub-cluster color proportional to their weight.

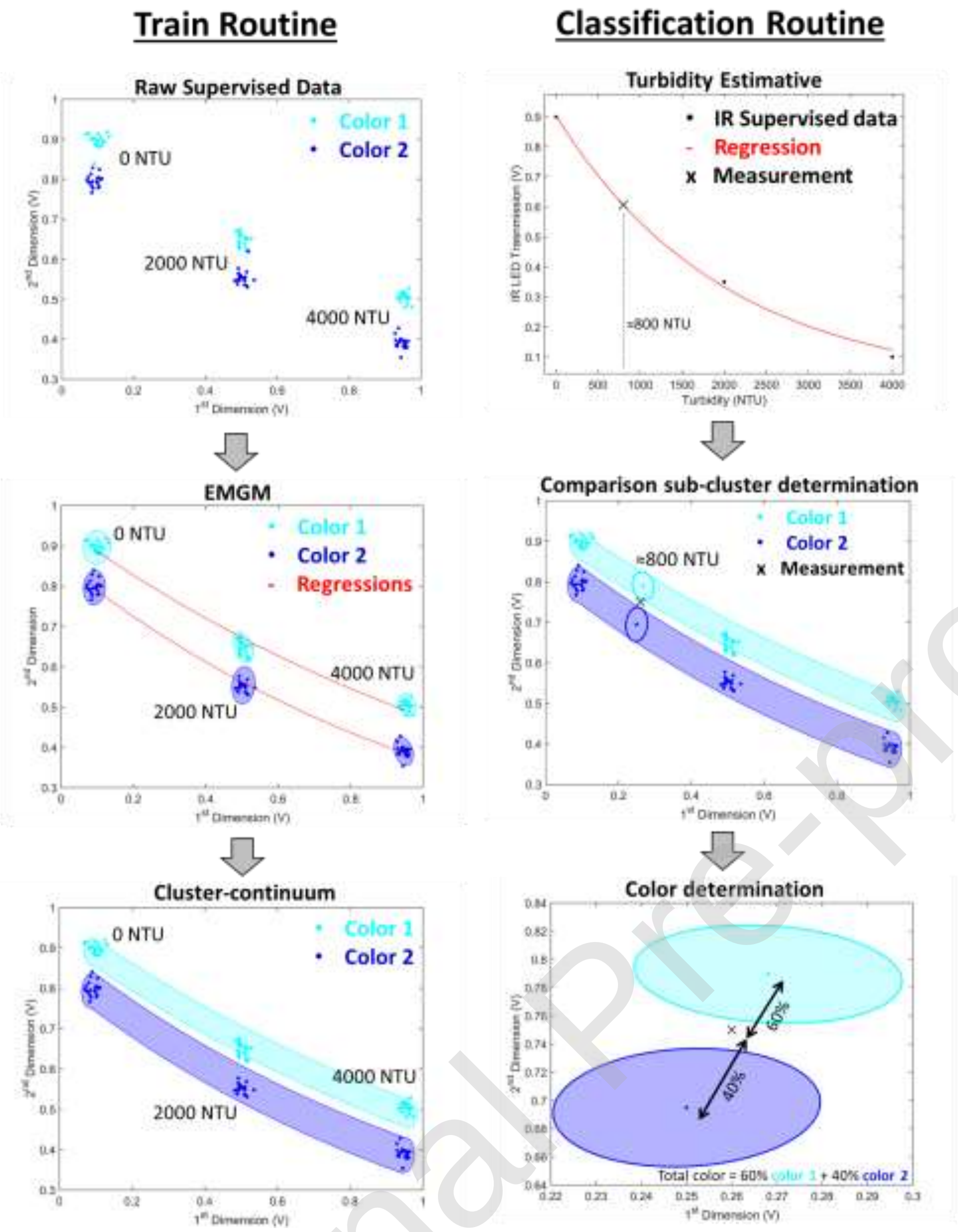


Fig. 6. Expectation maximization supervised Gaussian mixture algorithm scheme. Here an example using only the representation of 2 dimensions and 2 colors are presented. In the classification routine final step, the measurement point to be classified have the weight of the 2 sub clusters which makes its color a combination of the other 2.

4.3 Artificial neural networks

The last method is based on the application of the machine learning technique of artificial neural networks (ANN) [26]. ANN are computational models based on the biological operation and connections of the neurons of living beings. In the same way a neuron processes and transmits information to other neurons by synapses, the ANN will also have processing units that are connected between each other. Each processing unit is nonlinear and will have inputs that will be processed with simple activation functions producing this way a response as outputs. Linear, sigmoid and hyperbolic tangent are example of functions usually used. When an output is produced, it will be transmitted to the other neurons by “synapses”. All the “synapses” will have a weight associated to them that is multiplied by the output and deliver to the next neuron. The general structure of an ANN is made of three parts or layers as can be seen in Figure 7. The input layer is responsible for receiving the information to be determined. In the case of our sensor data, this layer will receive the 8-dimensional measurements. The hidden layers are composed of the neurons responsible to extract the patterns associated with the processed being analyzed. If it is a complex problem to be solve, more hidden layers will be needed. On the other hand, using a large number of hidden layers to solve simple problems could lead to overfitting. In our case only one hidden layer composed by 8 neurons will be used as it showed to have lower estimation error. The neurons of this layer will have sigmoids as processing functions. The output layer is also composed by neurons and are responsible to present the final output of the network. The neurons of the output layer used in this work are three and have linear functions since the output will be the three color RGB transmittance values. This method is also IR LED independent since it is not mandatory to have its information as an input, providing that the training routine does not use its information also.

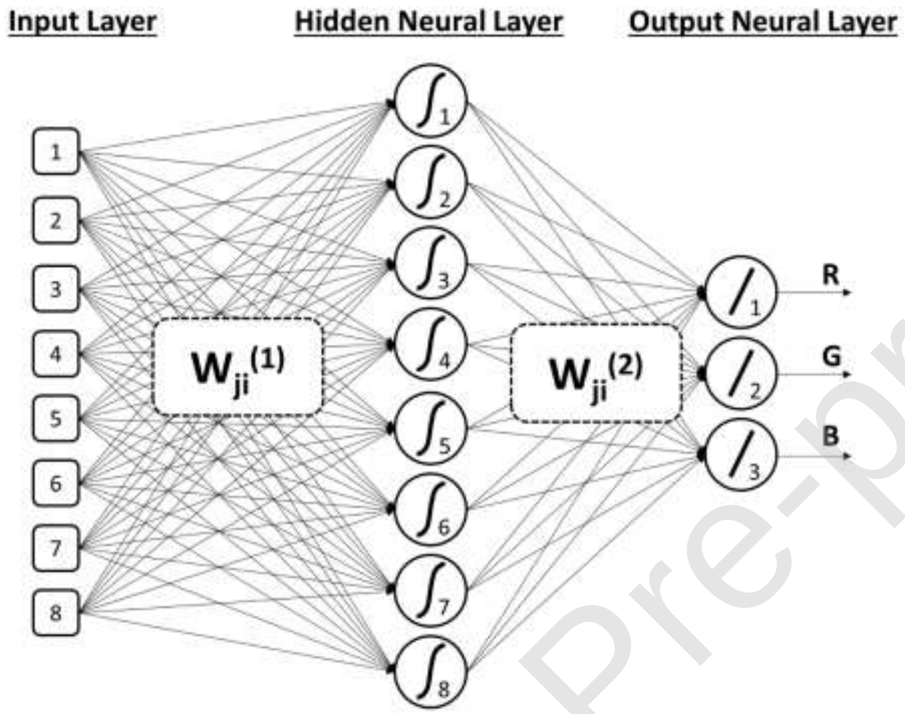


Fig. 7. General structure of the artificial neural network used in this work when IR LED information is also used. For the case of not using IR LED, the input layer will only have 6 entrees. Between each layer, a group of "synapses" denoted by a subscript number to *W* is presented. The final output will be the RGB color values to be determined.

The ANN needs to be trained with the labeled data to have the best final possible results. The most common training process and the one here applied consists in two distinct phases, the forward propagation and the back propagation. In the forward propagation we simply apply the set of weights of our network to the input data and calculate the final output. For the first forward propagation, the set of weights are selected randomly. This output values will then be used in the back propagation, were a measurement of the margin of error is made in respect to the expected labeled values and the weights will then be adjusted accordingly to decrease the final error. This is calculated generally by $weight = weight_{old} + rate \times error \times output$ The velocity of adjustment of the weights are controlled by a rate constant that was chosen to be 0.5 in this work. To validate the algorithm and prevent overfitting, 50% of the random selected training data was used for validation. An example of the evolution of the error rate with the number of epoch can be observed in Figure 8. With the increase of epochs, the training and validation error reach a point of stability with a minimal error rate value between them, indicating a good fit of the model.

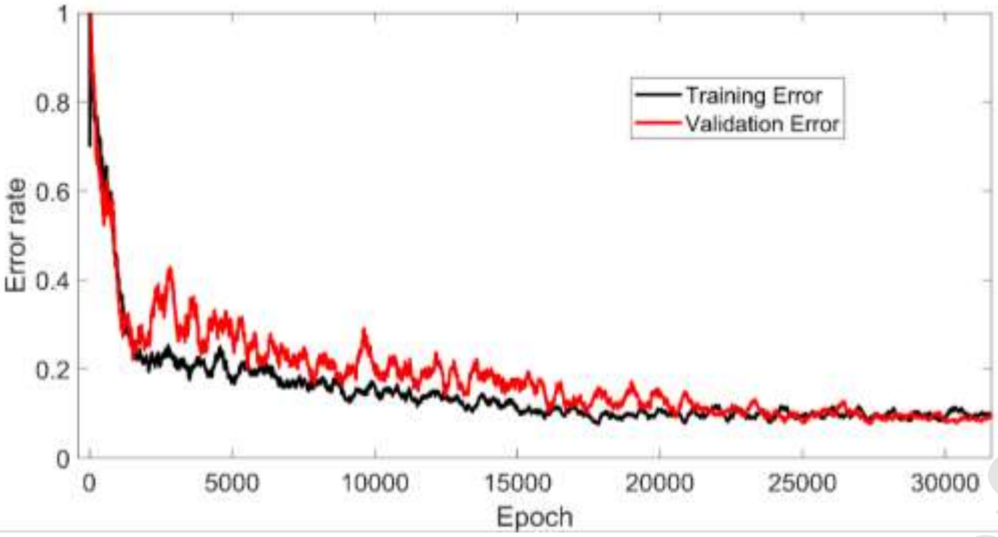


Fig. 8. ANN validation curves.

5. Color determination results

For each method discussed in 4, an analysis of the classification error of the color will be made with different percentages of trained samples. A random sampling cross validation method will be used [27] in which the samples that will be chosen for training are randomly selected. The samples that were left of training will be the ones that will have its color assessed. The process of train and classification will be performed over 100 cycles. For each cycle, the error ratio, the time of training and the time of classification will be calculated. The total error ratio and time related to each trained percentage will be calculated by the mean values. Two approaches will be made taking advantages of the proposed algorithms dependency of the IR LED information. The first approach will use the IR LED information to all the classifying methods and then an approach, without using this information, will be done to the algorithms that do not have a mandatory need of IR information, this is, EMGM and ANN. These processes were done in two programming languages, Matlab® R2014b and Python™ 2.7 with Numpy library, to assure that there is a proper comparison between the measured times within the different algorithms and independency with the programming language. A computer with an Intel® Core™ i5-2430M double core processor at 2.40 GHz and 6 GB of RAM was used having the 64-bit Windows® 10 as the operative system. The scripts were written for serial processing only and though only a single core was used.

5.1 Using the Infrared LED information

Although the use of IR LED information is not mandatory for the EMGM and ANN calculation algorithms, it is essential to the more direct approach of the regression models. The lines with dots in Figure 9 show the evolution of the error ratio with the increase of the trained data used for the regressions. As it can be seen, the error ratio stays practically in the same value (around 0.03) for 20% of the trained data and above. This was expected because of the error propagation associated with turbid media. Its natural Brownian behavior will create values with high scattering distribution independently of the optic system. Adding to this, the own sensor error (electric and optical fluctuations) and the error due to the fitted regression model will, together, define a lower limit error value for turbidity. Since the calculation of color has dependence on the value of turbidity, it will also achieve an error limit that seems to be reached with lower training data. Once this minimal error is achieved, it is impossible to go lower with this approach. To have a general idea of the performance of each method independently of the trained percentage, a calculation value was conducted based on the determination of the lower area of the error curves, just like an integration. The lower this value is, the higher the overall performance will be considered. For the regression models the value was 2.68 for Matlab and 3.08 for Python. A lower value was obtained for the EMGM algorithm which were 1.25 for Matlab and 1.52 for Python. The evolution of the error with the trained data percentage can be seen by circles marks in Figure 9. This can be attributed by the substantially lower value errors obtained (roughly an order of magnitude lower) for higher training percentages (40% and above) in comparison with the regression models approach. In contrast, for lower training percentages the error is higher. This proves that this method is excellent for measurements of liquids that have the same cyclic color variation in which good training data set can be obtainable and where the traditional method of color determination is outperformed. In relation to the ANN approach (with crosses in Figure 9), a nearly constant error value (around 0.05) was obtained from the 20% trained data to above, which shows that this could be the lowest error possible to achieve by this approach without overfitting. As expected, this higher error values will have influence on the overall performance values that will be the highest from all the methods, 4.15 and 4.54 for Matlab and Python respectively.

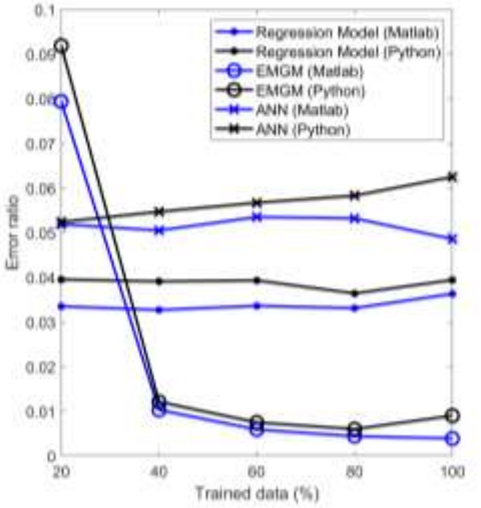


Fig. 9. Error ratio evolution with the trained data percentage (using IR information) for each classifier methods and for both Matlab and Python languages.

The overall time that the regression model approach needs to train the data is around 0.4 seconds for Matlab and 1 second for Pyhon, independently of the percentage of trained data used. In Figure 10a we see a representation of the training time of EMGM and ANN normalized to the time values obtained for the regression model training routine for each language for a better visual quality comparison. It is possible to conclude that the EMGM approach is more computational demanding than the ANN approach. The training process of EMGM is, at 20% training data, 65 times slower than the regression model method at the same percentage. When using 99% of the data for training it becomes 500 times slower at the same percentage. This is a linear increase and its easily predicted that the higher the quantity of data to be trained, the higher the time to train will be necessary. This is clearly a disadvantage compared to the other approaches. The ANN method, in turn, does not increase substantially its time to train when more data is available. It has a stable time that is around 30 times slower for Matlab and 100 times slower for Python. This differences between the two languages can be explained by the lack of optimization done to the Python code. The code was first written and optimized in Matlab and then directly adapted to Python without taking into account its particular differences as the examples of data structure that in Matlab is column-major while in the Numpy library of Python it is row-major.

While the training routine can be easily performed offline, the classification routine has higher importance due to the demand of high speeds for in-line real-time sensor monitoring and when using multiple sensors that retrieve data to the same platform for calculation. If the calculation algorithm is not fast enough, bottle neck processing effects will exist, and loss of data is a possibility. There are also applications in which high delay times are unacceptable, mainly if process automation is present. The same analysis performed to the training times was also done to the classifying times and can be seen in Figure 10b. The regression model approach time needed to classify a single measurement for Matlab was 0.006 seconds while for python was 0.0003 seconds. The EMGM method took about 10 times over the regression time to classify in Matlab, while in Python it was 60 times over, so it can be concluded that the EMGM method is in general a slow process. In contrast, the ANN method is very fast to classify which is expected because of the simple operation process needed. In Matlab it only took 0.03% of the time obtained from the regression models. In Python the time was 3 times slower, which again can be explained by the lack of optimization. Nevertheless, it is still in the same magnitude of the traditional regression models which is fast enough.

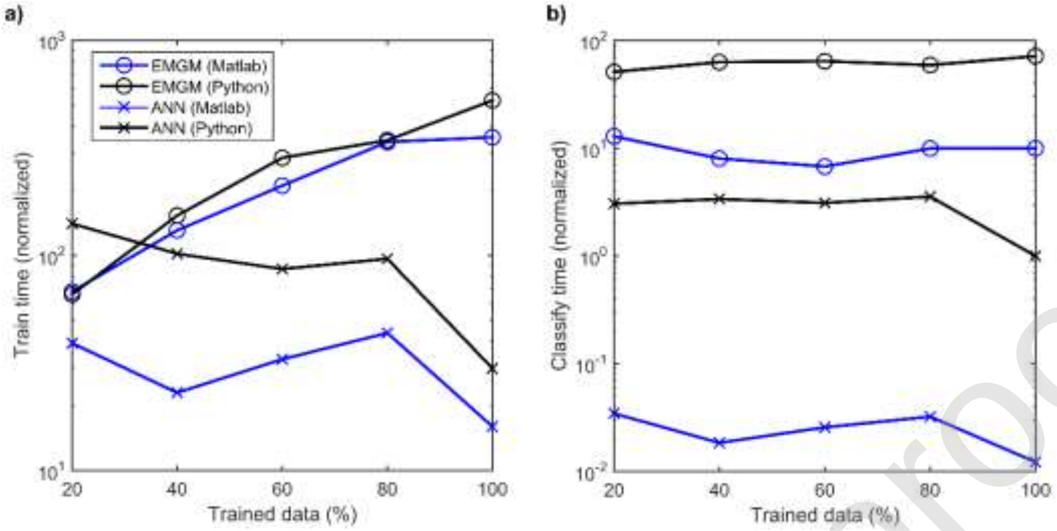


Fig. 10. a) Normalized training time (logarithmic scale) in relation to the regression model approach and its equivalent b) for the classify times, both with the variation of the training data percentage.

5.2 Without the Infrared LED information

As mentioned before, the infrared LED information is not mandatory for the implementation of the EMGM and ANN algorithms. While for the ANN approach the non-linear model is automatically determined by direct implementation to the labeled data, the same does not apply for the EMGM. This is important because the infrared LED information was directly used for a first estimate of the turbidity which, in turn, is used to calculate the color in EMGM. Without this information a recursive cycle that minimizes and compares the error of the expected turbidity for each individual dimension is done which will take more time to classify a measurement and is expected to increase the overall error because of its higher uncertainty, due to the reduction of information. In Figure 11 is represented the error ratio for both EMGM and ANN approaches with the different training data percentages and for the both programing languages.

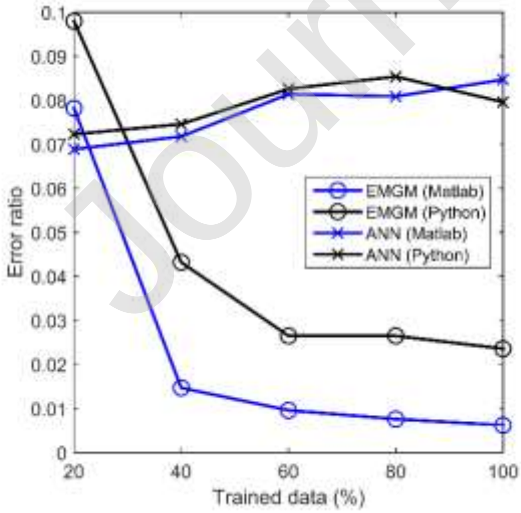


Fig. 11. Error ratio evolution with the trained data percentage (without IR information) for EMGM and ANN classifier methods and for both Matlab and Python languages.

As expected, the overall error obtained for both approaches without the IR information were higher than the ones obtained with the IR information. Nevertheless, the increase of error is not that higher that invalidates this approach, which can be very advantageous because it simplifies the necessary hardware of the sensor. Also, again, the EMGM performance was higher with the increase of trained data as observed before with the IR information. The overall performance value obtained for Matlab and Python were 1.48 and 3.14 respectively. The ANN approach has the same constant error behavior as observed before, having the overall performance of 6.21 and 6.37 for Matlab and Python. The normalized training times in relation to the previously regression models for the EMGM and ANN approach are represented in Figure 12a. As observed the EMGM had approximately the same increase time behavior with the increase of trained data while the ANN has a more constant value of 30 and 50 times slower for Matlab and Python respectively in relation to the regression models with IR. Because of the alternative away to estimate turbidity for the EMGM process, as mentioned before, the classification time was higher without the IR information than with it (3x slower). This not happened to the ANN process were the values stayed the same as with IR information. Figure 12b shows the normalized values of classification time.

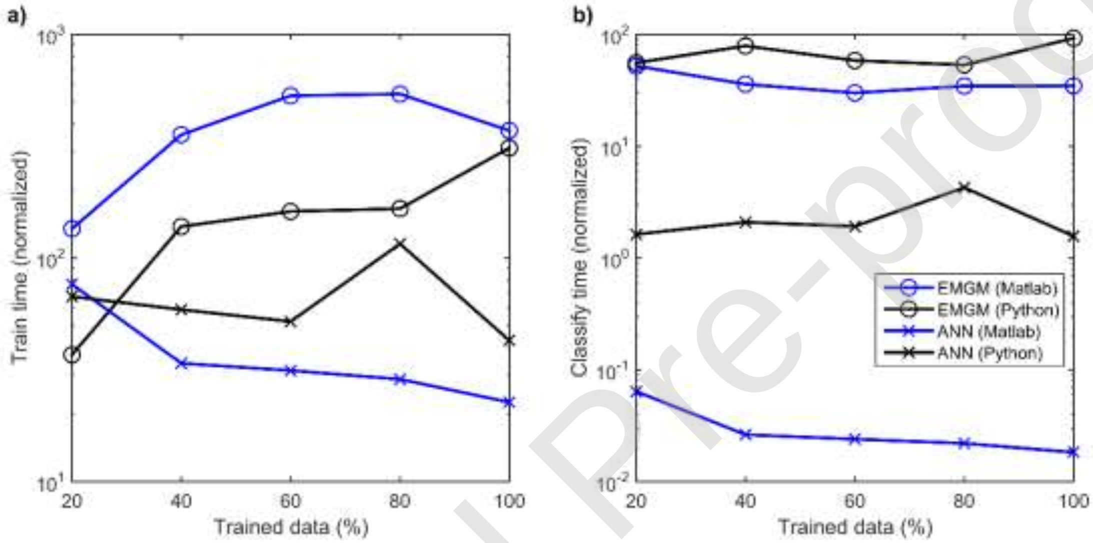


Fig. 2. a) Normalized training times (logarithmic scale) and the b) classify times for the ANN and EMGM methods without IR information, in relation to the regression model and with the variation of the training data percentage.

5.3 Resume table and final considerations

As seen above, all the approaches here presented, including the ones based in machine learning as a proof of concept, are feasible and each one has its advantages and disadvantages. The traditional approach of regression models presented advantages in the faster training times while maintaining a low error. Unfortunately, it needs the presence of the IR information which requires the more complex hardware. The ANN revealed to be as fast, if not fastest, to classify as the regression models. It as the great advantage of not needing to know the physical models and can therefore be directly applied to the labeled data without prior behavior knowledge. It also does not need the IR information like the EMGM approach. The EMGM has the disadvantage to be the slowest method of all and needs to know the physical models, but it will be highly precise if the objective color liquid to measure have a cyclic pattern where good data for training is obtainable. In Table 2 a resume of all the advantages and disadvantages observed can be seen.

Table 2
Resume table of the advantages and disadvantages of Regression model, EMGM and ANN color determination methods.

	Regression Models		**EMGM**		**ANN**	
	Matlab	Python	Matlab	Python	Matlab	Python
Physical model	✓		✓		✗	
Overall error ratio performance	2.68	3.08	1.25	1.52	4.15	4.54
Train speed	Fast		Very slow		Slow	
Classify speed	Fast		Very slow		Very Fast	Fast
Mandatory IR Information	✓		✗		✗	
Overall error ratio performance (No IR)	------		1.48	3.14	6.21	6.37
Observations	------		Particularly good for cyclic color variation patterns		Potentially good for real-time measurements	

6. Conclusion

In this paper a low-cost optic color sensor for turbid liquids was presented. Three different methods of data analysis were developed with each one having its advantages and disadvantages. The traditional regression model showed to be fast and with small error for standard or occasional measurements, but it needs IR LED to be able to measure color. Simpler and low-cost sensors, without the IR component, can be developed if EMGM or ANN is the methodology used. If the intended liquid to be measured has a very well-known range of colors and turbidities, where easily trained data can be obtained, then the EMGM proves to be very effective than the other methods with its low error ratio. The drink industry is an example of this category where quality control is essential for high precise measurements and where is expected a well-defined range of color values. If the intended propose is fast real-time measurement for a sensor network or using an internet of things (IoT) platform, the ANN has the potential to be very fast without compromising highly its error ratio value. ANN is also easily to implement since it does not need any previously known physical models.

For a more complete and comprehensive study to validate these preliminary conclusions, application of these algorithms to real day-to-day scenarios, like the ones mentioned before, will need to be performed. In these scenarios, multiple labeled samples and a large dataset can be obtained. Nevertheless, results here obtained, as a proof of concept, demonstrate the potentiality of the application of machine learning to multivariate data.

Author_statement

Daniel P. Duarte: Conceptualization, Methodology, Software, Formal analysis, Writing - Original Draft. **Rogério N. Nogueira**: Writing - Review & Editing. **Lúcia Bilro**: Validation, Writing - Review & Editing, Supervision, Funding acquisition.

Funding

This work is funded by FCT/MEC through national funds and when applicable co-funded by FEDER – PT2020 partnership agreement under the project INITIATE-IF/FCT-IF/01664/2014/CP1257/CT0002 and UIDB/50008/2020 and PhD fellowship (Daniel Duarte SFRH/BD/130966/2017).

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Z. Fang, M. Zhang, Y. Sun, J. Sun, How To Improve Bayberry (Myrica rubra Sieb. et Zucc.) Juice Color Quality: Effect of Juice Processing on Bayberry Anthocyanins and Polyphenolics, J. Agric. Food Chem. 54 (2006) 99–106. doi:10.1021/jf051943o.

[2] T.R. Burns, J.P. Osborne, Loss of Pinot noir Wine Color and Polymeric Pigment after Malolactic Fermentation and Potential Causes, Am. J. Enol. Vitic. 66 (2015) 130–137. doi:10.5344/ajev.2014.14061.

[3] B. Gandul-Rojas, L. Gallardo-Guerrero, M. Roca, R. Aparicio-Ruiz, Chromatographic Methodologies: Compounds for Olive Oil Color Issues, in: R. Aparicio, J. Harwood (Eds.), Handb. Olive Oil Anal. Prop., Springer US, Boston, MA, 2013: pp. 219–254. doi:10.1007/978-1-4614-7777-8.

[4] C. Pasquini, Near infrared spectroscopy: A mature analytical technique with new perspectives – A review, Anal. Chim. Acta. 1026 (2018) 8–36. doi:10.1016/j.aca.2018.04.004.

[5] A.F. Omar, M.Z. MatJafri, Water Quality Measurement using Transmittance and 90° Scattering Techniques through Optical Fiber Sensor, in: 2008 6th Natl. Conf. Telecommun. Technol. 2008 2nd Malaysia Conf. Photonics, IEEE, 2008: pp. 17–21. doi:10.1109/NCTT.2008.4814227.

[6] L. Bilro, S.A. Prats, J.L. Pinto, J.J. Keizer, R.N. Nogueira, Design and performance assessment of a plastic optical fibre-based sensor for measuring water turbidity, Meas. Sci. Technol. 21 (2010) 107001. doi:10.1088/0957-0233/21/10/107001.

[7] M.A.P. Garcia, R.M. Vega, C. de la Torre Fernandez, J.A.B. de la Fuente, L.M.C. Carcel, Full-range, true on-line turbidimeter based upon optical fibers for application in the wine industry, in: 2008 IEEE Instrum. Meas. Technol. Conf., IEEE, 2008: pp. 130–134. doi:10.1109/IMTC.2008.4547017.

[8] R. Crespo, L.M. Cárcel, M.A. Pérez, I. Nevares, M. del Álamo, Suitable at-line turbidity sensor for wine fermentation supervision, in: Int. Conf. Food Innov., Valencia, 2010: pp. 1–4.

[9] S. Yeoh, M.Z. Matjafri, K.N. Mutter, A.A. Oglat, Plastic fiber evanescent sensor in measurement of turbidity, Sensors Actuators, A Phys. 285 (2019) 1–7. doi:10.1016/j.sna.2018.10.042.

[10] I. Hussain, K. Ahamad, P. Nath, Water turbidity sensing using a smartphone, RSC Adv. 6 (2016) 22374–22382. doi:10.1039/C6RA02483A.

[11] F. Jiménez-Márquez, J. Vázquez, J. Úbeda, J.L. Sánchez-Rojas, High-precision optoelectronic sensor device for monitoring fermentation kinetics and maceration of wine, in: U. Schmid, J.L. Sánchez de Rojas Aldavero, M. Leester-Schaedel (Eds.), Proc. SPIE - Int. Soc. Opt. Eng., 2013: p. 87630I. doi:10.1117/12.2016935.

[12] F. Jiménez-Márquez, J. Vázquez, J.L. Sánchez-Rojas, Optoelectronic sensor device for monitoring the maceration of red wine: Design issues and validation, Measurement. 63 (2015) 128–136. doi:10.1016/j.measurement.2014.12.009.

[13] C. Novo, L. Bilro, R. Ferreira, N. Alberto, P. Antunes, R. Nogueira, J.L. Pinto, Optical fibre monitoring of Madeira wine estufagem process, in: M.F.P.C. Martins Costa (Ed.), Proc. SPIE - Int. Soc. Opt. Eng., 2013: p. 8785EX. doi:10.1117/12.2025397.

[14] S. Sumriddetchkajorn, K. Chaitavon, Y. Intaravanne, Mobile-platform based colorimeter for monitoring chlorine concentration in water, Sensors Actuators B Chem. 191 (2014) 561–566. doi:10.1016/j.snb.2013.10.024.

[15] C.M. Bishop, Mixture Models and EM, in: M.I. Jordan, J. Kleinberg, B. Schölkopf (Eds.), Pattern Recognit. Mach. Learn., 1st ed., Springer-Verlag New York, 2007: pp. 423–459.

[16] D. Duarte, R.N. Nogueira, L.B. Bilro, Semi-supervised gaussian and t-distribution hybrid mixture model for water leak detection, Meas. Sci. Technol. (2019). doi:10.1088/1361-6501/ab3b48.

[17] N. Ding, H. Ma, H. Gao, Y. Ma, G. Tan, Real-time anomaly detection based on long short-Term memory and Gaussian Mixture Model, Comput. Electr. Eng. 79 (2019) 106458. doi:10.1016/j.compeleceng.2019.106458.

[18] A. Ukil, J. Bernasconi, Neural Network-Based Active Learning in Multivariate Calibration, IEEE Trans. Syst. Man, Cybern. Part C (Applications Rev. 42 (2012) 1763–1771. doi:10.1109/TSMCC.2012.2220963.

[19] W. Cao, X. Wang, Z. Ming, J. Gao, A review on neural networks with random weights, Neurocomputing. 275 (2018) 278–287. doi:10.1016/j.neucom.2017.08.040.

[20] N. Oliveira, D. Duarte, F. Gonçalves, P. Costa, S. Vieira, N. Fontes, P. Prior, A. Graça, M. Figueira, L. Bilro, R. Nogueira, Winegrid ® : the remote and real-time wine production process monitoring system, in: 40th World Congr. Vine Wine, Sofia, Bulgaria, 2017: pp. 1–6.

[21] C. Novo, L. Bilro, N. Alberto, P. Antunes, R. Nogueira, J.L.L. Pinto, Plastic optical fibre sensor for Madeira wine monitoring, in: M.F.P.C. Martins Costa, R.N. Nogueira (Eds.), Proc. SPIE - Int. Soc. Opt. Eng., 2014: p. 92862Q. doi:10.1117/12.2060322.

[22] P. Ribéreau-Gayon, Y. Glories, A. Maujean, D. Dubourdieu, Handbook of Enology, John Wiley & Sons, Ltd, Chichester, UK, 2006. doi:10.1002/0470010398.

[23] Libelium Comunicaciones Distribuidas S.L, Smart Water Technical Guide, V.7.5. (2019) 94. http://www.libelium.com/downloads/documentation/smart_water_sensor_board.pdf (accessed September 5, 2019).
[24] H. Tai, D. Li, C.C.C.C. Wang, Q. Ding, C.C.C.C. Wang, S. Liu, Design and characterization of a smart turbidity transducer for distributed measurement system, Sensors Actuators A Phys. 175 (2012) 1–8. doi:10.1016/j.sna.2011.11.028.
[25] C.F. Bohren, D.R. Huffman, Absorption and scattering of light by small particles, 1983. doi:10.1038/ncomms1111.
[26] I.N. da Silva, D. Hernane Spatti, R. Andrade Flauzino, L.H.B. Liboni, S.F. dos Reis Alves, Artificial Neural Networks, Springer International Publishing, 2017. doi:10.1007/978-3-319-43162-8.
[27] S.B. Kotsiantis, Supervised machine learning: A review of classification techniques, Informatica. 31 (2007) 249–268.

Daniel P. Duarte is presently a PhD candidate at Department of physics in Aveiro University, Portugal and researcher at Telecommunications Institute Aveiro, Portugal. He received the M.Sc. degree in engineering physics from the same university in 2013. His currently research interests are the application of machine learning and data fusion algorithm to optical sensor data analysis and development of low-cost fiber optic-based sensors.

Rogério N. Nogueira is presently a Principal Research Scientist at Telecommunications Institute (IT) in Portugal, Co-founder and CEO of Watgrid, SA and co-founder and vice-president of the Portuguese Optical Society. Prior to that he was an invited Professor at Aveiro University and Innovation Manager at Nokia Siemens Networks. He received the PhD degrees from the Aveiro University in 2005 and an Executive Certificate on Management and Leadership from MIT. He coordinates a group at IT in the topics of Fiber Optical Components and Devices for Optical Communication and Sensors.

Lúcia Bilro is presently a Researcher at Telecommunications Institute (IT) in Portugal, Co-founder and CTO of Watgrid, SA and Board Member of the Portuguese Optical Society (SPO). Prior to that she was Post-doc research fellow at IT. She received the PhD degree from the Aveiro University in 2011. Her currently research interests are Bragg sensors, plastic optical fiber sensors, physics medicine and rehabilitation, environmental monitoring.