

Fusion of Online Assessment Methods for Gynecological Examination Training: a Feasibility Study

E.A.M.G. SOARES¹ and R.M. MORAES²

Received on June 10, 2017 / Accepted on April 12, 2018

ABSTRACT. The objective of this paper was to determine if a fusion of online assessment methods is a feasible methodology for online assessment of performance of users inside virtual reality simulators. Three different forms of the Fuzzy Naive Bayes method based on statistical distributions were used to assess specific tasks and the fusion of information was performed by a Weighted Majority Voting system. Data was compiled representing a portion of the Gynecological Examination, which is a checkup examination that is routinely performed for women and is paramount in finding earlier cases of cervical cancer. Confusion matrices and Kappa coefficients were obtained using a Monte Carlo simulation for this method. From the analysis of these results, it is possible to confirm that this method performed well, with a substantial agreement degree.

Keywords: Fusion of assessment methods, online assessment, Fuzzy Naive Bayes, virtual reality, gynecological examination.

1 INTRODUCTION

It is well known that the more a given task is performed, the more practicality and expertise will be achieved. For some areas, especially in medicine, the lack of practice in certain procedures can have consequences ranging from simple complications to the patient death. Tools have been created to aid in the learning and enhancement of certain skills, considering that practice in medicine is of vital importance. For the health sciences, the most popular method for training is the use of guinea pigs, corpses and mannequins, but these have limitations such as wear of the material over time and lack of representation of the real characteristics of a human being. Another method used in medical-schools is allowing students to practice with real cases under the supervision of a physician, which limits their training to simple often-occurring cases, causing sometimes discomfort to the patient [17].

*Corresponding author: Elaine Anita de Melo Gomes Soares – E-mail: elaineanita1@gmail.com

¹Graduate Program in Decision Models and Health, Universidade Federal da Paraíba João Pessoa, PB, Brasil. E-mail: elaineanita1@gmail.com

²Departamento de Estatística, Universidade Federal da Paraíba João Pessoa, PB, Brasil. E-mail: ronei@de.ufpb.br

A solution proposed in 1999 [1], which has been improved since then, is the use of virtual reality (VR) simulators for the training of certain medical procedures. Training systems implemented using virtual reality have been used in several areas [2] and its main purpose is to produce the sensation of immersion for the user in order to make the training of the chosen procedure executed as realistic as possible. Attached to the VR system, it is possible to have one or more assessment systems with the function of analyzing the data generated by the execution of the procedure, and returning a report to the user informing them about their performance. Additionally, it has been proven that surgeons trained in virtual reality systems can obtain better results when compared to those trained by traditional methods [10].

In order to be able to return information about the user's performance, it is important to analyze the user's actions in this environment. Information about the user's movement in three-dimensional space can be captured through common peripherals, such as mouse and keyboard, but it is possible to use more specific haptic devices, which return information such as forces and angles. There are several ways to analyze the information collected during the procedure. These are classified as offline or online when related to the speed in which the information is returned.

The offline assessment is characterized by recording the procedure for further analysis by a professional in the area, who generates a report and returns it to the user. Examples of applications of this type can be found in the literature [1] [16] [24]. The online assessment monitors the user's actions to gather data, such as angle, force, among others, and then compares them with performance classes previously defined by a specialist in this procedure. After the procedure is finalized, the result of this comparison is returned to the user in a maximum time of one second.

A recurring problem is that a few moments after the simulated procedure is done, the user cannot clearly remember the exact movements they performed, thus reducing the learning [18]. The solution for that problem lies on the online assessment. Since it is incorporated into the simulator, the result of the simulation is returned as soon as the simulation is completed, within a range of less than one second, thereby increasing the amount of information captured by the user [15]. This is the main feature that makes the online approach more suitable for amplifying the user's learning when compared to offline, since the user can identify their errors and correct them at the next execution.

These assessment systems can be based on logic, probabilistic models, fuzzy models, neural networks, or mixture models, thus creating hybrid systems. In the area of health, several training systems have been proposed. Some of these use machine learning, fuzzy sets, or Naive Bayes methods and variations [8] [11] [15].

As mentioned before, there are several methods in the literature that reach this result, but it has already been proven that certain methods have better performance when applied to certain types of data, thus the relevance of the statistical distribution of the data. Traditionally, only one method is used for each application, but proposals have already been made in which more than one method is used, thus composing a fusion of assessment systems [26]. By using the fusion of information,

every piece of data will be analyzed by a method that considers its statistical distribution, which may lead to a more accurate result when compared to a single method assessing them all.

The purpose of this article is to analyze the feasibility of fusing specific methods for assessment of users' performance in a virtual reality training system for a simulation of Gynecological Examination. More precisely, we will use variations of the Fuzzy Naive Bayes [17] method to assess certain variables of the exam and computational granularity for the task of fusing the results of these methods.

In the coming sections, important concepts for the understanding of the presented problem will be explored, as well as the bibliographic review mentioned previously. Then, the methodology used in order to achieve the objective of this proposal will be presented. Lastly, this paper will finish with the discussion of the results and then the conclusion.

2 SELECTED FUNDAMENTALS

2.1 Gynecological Examination

Gynecological Examination is a procedure that aims to identify cervical cancer and lesions related to the Herpes and HPV viruses. This examination is of utmost importance for women's health because, in addition to allowing the treatment of HPV and Herpes, it helps to identify cervical cancer in its early stages. It was estimated in 2009 that there is a reduction of 80% of cervical cancer mortality when it is found in women 25 to 65 years old and treatment is performed [5]. In addition, according to the INCA (Instituto Nacional de Cancer), from 2010 to 2014, there were around twenty-six thousand deaths in the world related to cervical cancer [12]. INCA is the national cancer institute that aims to prevent and treat cancer in Brazil. In addition to its social work, this institute also contributes to the society by collecting data and carrying out studies about cancer.

This examination consists of the following steps: anamnesis, breast examination, examination of the abdomen and examination of the external and internal genitals. Anamnesis is the collection of information about the patient, such as age, sex, number of children, etc. The next four steps involve physical contact with the patient, but this work will focus only on the last stage of the examination, given limitations of technology. This is the examination of the internal genitalia, firstly observing the external part looking for anomalies on the distribution of pubic hair and deformations on the patient's lower lips and then inserting the speculum to locate any wounds on the vaginal walls and analyze the cervix. In this part, it is necessary to detect abnormal characteristics and, from these, the prescription of exams according to the diagnosis of the doctor. Finally, it is necessary to collect material for cytological, bacteriological and cervical mucus analysis using the Ayre spatula and a Cytobrush, and characteristics such as elasticity, roughness and presence of tumors should be observed through touch [3].

For this paper, some of the variables presented in the phase of the exam described above will be simulated and analyzed. Additionally, in order to assess the performance of the user, parameters

will be used for a healthy patient. Table 1 presents the variables and their respective data distributions, which were used to choose the assessment method for each variable. The insertion of the speculum has an acceptable variation around its entrance angle, in which the patient is not harmed, i. e., this insertion has a variation in relation to its central point of entrance. This problem can be modeled through a Gaussian density probability, which mean is the central point of entrance and the standard deviation is the acceptance around it. The mean of the total time spent on each phase of the Gynecological Examination is modeled similarly to the lifetime of a process or product. The Exponential distribution is normally used on this context, which each parameter is the mean time spent in each phase [23]. The identification or not a determined anomaly follows the Bernoulli distribution, in which only two results are possible. The cervix area was divided in 8 sectors and the final event was having covered or not each sector. As each sector is independent with respect to others, the Binomial distribution is used to model all area. The Binomial distribution is a general case of the Bernoulli distribution, and for this reason these two last variables were modeled as Binomial distributions [25].

Table 1: Variables used to user assessment in the VR simulation.

Variable	Statistic Distribution
1. Anomaly in the Distribution of Pubic Hair	Bernoulli
2. Anomaly in the Lower Lips' Structure	Bernoulli
3. Total Time Examining the External Genitalia	Exponential
4. Angle of Input on the Insertion of the Speculum	Gaussian
5. Total Time Inserting the Speculum	Exponential
6. Cervix Area Covered with the Ayre Spatula	Binomial
7. Cervix Area Covered with the Cytobrush	Binomial
8. Total Time with the Ayre Spatula	Exponential
9. Total Time with the Cytobrush	Exponential

2.2 Fusion of Assessment Methods

The fusion of assessment methods can be accomplished in several different ways. From the analysis of individual results to the modifications in the calculations of each method [26], the fusion aims at the use of different assessment methods to improve the task of assessing the performance of the user.

Techniques for merging methods have been studied since 1990 [29]. There are three different types of aggregators [26]. The first works before the results are generated, that is, in the body of the method itself. The second starts after the methods inform their results. The third is specialized for methods which results are fuzzy, that is, number in the interval $[0, 1]$. Three different types of fusion were found for the first group described above. These are the dynamic selection of method, the grouping and structuring of methods and, finally, the hierarchical mixture of experts.

For the second group, there are two techniques. These are the voting method [13] and the behavior-knowledge space method. As the name implies, the voting method adds up the number of times each class appears in the results, selecting the most voted class. There is also a way to organize the classes by an order of precedence. For this, there are methods that reduce or reorder classes into groups. For the reduction, methods of union or neighborhood intersection are used. In addition, for the reordering, methods of class precedence, of class with greater relevance and of logistic regression are used.

For the third group, that is, methods returning fuzzy measures, Bayesian fusion methods are used, which may be a simple Bayesian mean or a Bayesian integration, fuzzy integral, Dempster-Shaffer combination, fuzzy templates, product of experts, or neural networks. The fuzzy integrals may be from Sugeno, Choquet, or Weber.

In this paper, the results of each method will be fused, i.e., the fusion will occur after all assessment methods have processed their variables. To make this process simpler from the problematic and computational points of view, only a variety of methods from the one presented in the next topic will be used. The input and the method will vary according to the distribution of the data and the output will always be a label or degree of membership of the data for the performance class.

2.3 Fuzzy Sets and Probability

Let there be a space of objects X with a generic element x . Given that, $X = \{x\}$. A fuzzy set A can be defined in X characterized by a membership function $\mu_A(x)$ which correlates each point x in X to a real number in the interval $[0, 1]$. The value of $\mu_A(x)$ represents the degree of membership of x in A [30]. For example, if $\mu_A(x_0) = 0$, it is said that x_0 does not belong to A ; if $\mu_A(x_1) = 1$, it is said that x_1 belongs to A ; and if $\mu_A(x_2) = 0.7$, it is said that the membership degree of x_2 in A is 0.7.

Furthermore, a fuzzy set A with membership function $\mu_A(x)$ can be expressed by the set of its α -cuts. Then, it is denoted by A_α and the following is true:

$$A_\alpha = \{x \in X | \mu_A(x) \geq \alpha\} \quad (2.1)$$

The membership function $\mu_A(x)$ can also be represented in terms of its α -cuts[6]:

$$\mu_\alpha = \sup_{\alpha \in [0,1]} \min\{\alpha, \mu_{A_\alpha}(x)\} \quad (2.2)$$

In 1968, Zadeh introduced the concept of probability for fuzzy events [31]. Let B be a σ -field of Borel subsets in R^n and P be a probability measure over Ω . Let F be a fuzzy event in B with pertinence function $\mu_F : R^n \rightarrow [0, 1]$. the probability of F is defined by the integral of Lebesgue-Stieljes:

$$P(F) = \int_{F \subseteq R^n} \mu_F(x) dP = E(\mu_F)$$

i.e., the probability of a fuzzy event F is the mathematical expectation of its membership function. It can be rewritten as:

$$P(F) = \int_{F \subseteq R^n} \mu_F(x)P(x)dP$$

Some fuzzy versions for the Naive Bayes classifier were proposed and in this work we follow the version proposed by Störr [28], which uses the concept of probability introduced here and was used by [17] as a kernel of an assessment system for training based on VR.

2.4 Fuzzy Naive Bayes

Formally, let there be the classes of performance in space of decision $\Omega = \{1, \dots, M\}$ where M is the total number of classes of performance. Let there be $\omega_i, i \in \Omega$ the class of performance for a trainee. It is possible to determine the class of performance most probable for this trainee given a data vector $X = \{X_1, X_2, \dots, X_n\}$ and it is assumed each $X_k, k = 1, \dots, n$, is a fuzzy variable, with normalized membership functions $\mu_i(X_k)$, where $i = 1, \dots, M$. The method is defined by [17]:

$$P(w_i|X) = \frac{P(w_i)}{S} * \prod_{k=1}^n P(X_k|W_i) * \mu_i(X_k) \quad i \in \Omega \tag{2.3}$$

where S is a scale factor. In order to reduce this method’s computational complexity, the logarithm function was applied to Equation 2.3, replacing multiplications by additions. Thus, rewriting $P(w_i|X)$ as $g(w_i, X_1, \dots, X_n)$, given by

$$g(w_i, X_1, \dots, X_n) = \ln[P(w_i)] + \ln(1/S) + \sum_{k=1}^n \{ \ln[P(X_k|W_i)] + \ln[\mu_i(X_k)] \} \tag{2.4}$$

The classification rule for Fuzzy Naive Bayes is:

select performance class w_i for the vector X if

$$g(w_i, X_1, \dots, X_n) > g(w_j, X_1, \dots, X_n) \quad \text{for all } i \neq j \quad i, j \in \Omega \tag{2.5}$$

Although this method is very useful for assessment tasks, it does not assume a specific distribution [22]. In this paper, three different variations of the method presented above were used. These are Fuzzy Exponential Naive Bayes (FExpNB) [22], Fuzzy Gaussian Naive Bayes (FGauNB) [19], and Fuzzy Binomial Naive Bayes (FBinNB) [21]. For the Fuzzy Exponential Naive Bayes method, the $\ln[P(X_k|W_i)]$ element from Equation 2.3 is given by [22]:

$$\ln[P(X_k|W_i)] = \ln[\lambda_{ki}e^{-\lambda_{ki}X_k}] = \ln[\lambda_{ki}] - (\lambda_{ki} * X_k)$$

where λ_{ki} is the inverse of the mean of the variable X_k learned from the training data. For the Fuzzy Gaussian Naive Bayes method, the $\ln[P(X_k|W_i)]$ element from Equation 2.3 is given by [19]:

$$\ln[P(X_k|W_i)] = \ln \left[\frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(X_k - \mu_k)^2}{2\sigma_k^2}} \right]$$

$$\ln[P(X_k|W_i)] = \ln(1/\sigma_k) - \frac{(X_k - \mu_k)^2}{2\sigma_k^2}$$

where μ_k is the mean and σ_k is the standard deviation learned from the training data.

For the Fuzzy Binomial Naive Bayes method, the $\ln[P(X_k|W_i)]$ element from Equation 2.3 is given by [21]:

$$\ln[P(X_k|W_i)] = \ln \left[\binom{\eta_k}{X_k} p_{ki}^{X_k} (1 - p_{ki})^{(\eta_k - X_k)} \right]$$

$$\ln[P(X_k|W_i)] = \ln(\eta_k!) - [\ln(X_k!) + \ln(\eta_k - X_k)!] + X_k \ln(p_{ki}) + (\eta_k - X_k) * \ln(1 - p_{ki})$$

where η_k is the number of experiments observed for the variable X_k and p_{ki} is the success probability, both learned from the training data.

It can be observed that each variable was assessed by the method corresponding to their statistical data. Table 2 presents the different types of distribution from database used here and their respective assessment methods.

Table 2: Statistical distributions and their respective assessment methods.

Statistical Distribution	Assessment Method
Exponential	Fuzzy Exponential Naive Bayes
Gaussian	Fuzzy Gaussian Naive Bayes
Binomial	Fuzzy Binomial Naive Bayes
Bernoulli	Fuzzy Binomial Naive Bayes

2.5 Weighted Majority Voting

The majority rule, or voting system, is a very popular aggregation method used in many different cases. What makes this method so popular is its simplicity, which makes processing faster and less complex. There are many variations of this method, some more complex use trees or fuzzy sets in its calculations. However, the one used in this paper is the weighted version of the most traditional approach [13]. The label outputs can be represented as votes of support for the classes of performance as

$$d(D_i(X), w_j) = \begin{cases} 1, & \text{if } D_i(X) = w_j \\ 0, & \text{otherwise} \end{cases}$$

where $d(D_i(X), w_j)$ is the vote of the assessment of the data X with the assessment method D_i for the performance class w_j .

The final decision for the data vector X is obtained through weighted voting as

$$h_j(X) = \sum_{i=1}^n b_i d(D_i(X), w_j)$$

where b_i is the given weight for assessment method D_i .

In this study, this decision system was applied as an aggregator on the fusing of the output from the mentioned assessment methods. It sums all the votes for each performance class and the class that presents the higher amount of votes is the one selected. In order to make this voting more realistic, weights are assigned for each assessment method depending on the relevance of their respective variables to the simulation. The weights were assigned by a specialist in the area.

2.6 Confusion Matrix

The confusion matrix is a table used to measure the performance of an assessment method on a data set for which there exists an expected answer. A simple way of measuring the percentage of correct decisions made by the method is to compute the sum of the values in the main diagonal of the matrix divided by the sum of all values of the matrix [9]. The following table shows the confusion matrix for a three class assessment system.

Table 3: Example of a confusion matrix.

Predicted as - >	C1	C2	C3
C1 = very good	X_{11}	X_{12}	X_{13}
C2 = need training	X_{21}	X_{22}	X_{23}
C3 = unacceptable	X_{31}	X_{32}	X_{33}

2.7 Kappa Coefficient

The Kappa Coefficient K is widely used in the literature of pattern classification [7]. This coefficient was proposed by Cohen [4] and it is a weighted measure which takes into account agreements and disagreements between two sources of information. From a confusion matrix:

$$K = \frac{P_0 - P_c}{1 - P_c} \tag{2.6}$$

with P_0 and P_c as:

$$P_0 = \frac{\sum_{i=1}^M n_{ii}}{N} \quad \text{and} \quad P_c = \frac{\sum_{i=1}^M n_{i+} n_{+i}}{N^2} \tag{2.7}$$

where n_{ii} is the total of the main diagonal, n_{i+} is the total of line i , n_{+i} is the total of column i , M is the total number of classes, and N is the total of possible decisions in the classification matrix.

The variance of the Kappa Coefficient K , denoted by σ_K^2 is described by [20] as:

$$\sigma_K^2 = \frac{P_0(1 - P_0)}{N(1 - P_c)^2} + \frac{2(1 - P_0) + 2P_0P_c - \theta_1}{N(1 - P_c^3)} + \frac{(1 - P_0)^2\theta_2 - 4P_c^2}{N(1 - P_c)^4} \tag{2.8}$$

where θ_1 and θ_2 are given by:

$$\theta_1 = \frac{\sum_{i=1}^M n_{ii}(n_{i+} + n_{+i})}{N^2} \quad \text{and} \quad \theta_2 = \frac{\sum_{i=1}^M n_{ii}(n_{i+} + n_{+i})^2}{N^3} \tag{2.9}$$

All Kappa coefficients and respective variances were computed, which are presented in the Results section of this paper. Additionally, according to Landis and Koch nomenclature [14], the Kappa coefficient can be interpreted as presented in the Table 4.

Table 4: Interpretation of Kappa Coefficient [14]

Kappa Coefficient	Agreement Degree
< 0.0	Poor
$0.00 \text{ } - 0.20$	Slight
$0.20 \text{ } - 0.40$	Fair
$0.40 \text{ } - 0.60$	Moderate
$0.60 \text{ } - 0.80$	Substantial
$0.80 \text{ } - \text{ } 1.00$	Almost Perfect

2.8 Simulation

It has been proven that several classification methods found in the literature can obtain better performance when applied to data from specific statistical distributions [27]. In order to test if the approach proposed in this paper is feasible, a Monte Carlo simulated data was generated based on the variables presented on Table 1. One set of data containing 100 samples per class was generated for both the training and the assessment tasks. In fact, 150 samples were generated, but the first 50 were discarded in order to prevent unwanted oscillations in the probability distributions.

Each sample had three classes of performance, which then had 9 dimensions, one for each variable displayed on Table 1. Each variable was analyzed by their assigned assessment method and their results stored to then be fused using the weighted majority voting system. The weights used for each assessment method are displayed on Table 5. These weights were determined by the relevance of their variables to the procedure simulated by this work.

Table 5: Weights for the fusion of the assessment methods' output.

	FGauNB	FBinNB	FExpNB
Weight	2	3	1

The parameters used to generate the data for each class of performance and dimension are displayed on the tables 6 to 8. They were grouped on these tables by their data distribution. All distributions had their data generated using the same methodology, but different parameters, which were designed to be as similar to the real procedure as possible. Additionally, variables one and two were simulated as Binomial distributions with $\eta = 1$. The classes are considered as class one (C1) for very good performance, class two (C2) for acceptable performance and class three (C3) for unacceptable performance. The final result of the whole assessment process was stored in a confusion matrix, which will be presented in the results section.

Table 6: Parameters for variables one, two, six and seven.

	C1	C2	C3
η_1	1	1	1
p_1	0.1	0.6	0.9
η_2	1	1	1
p_2	0.2	0.6	0.9
η_6	8	8	8
p_6	0.9	0.70	0.5
η_7	8	8	8
p_7	0.95	0.8	0.5

Table 7: Parameters for variables three, five, eight, and nine.

	C1	C2	C3
λ_3	120 s	240 s	480 s
λ_5	60 s	120 s	360 s
λ_8	10 s	30 s	60 s
λ_9	20 s	40 s	80 s

Table 8: Parameters for variable four.

	C1	C2	C3
μ_4	45°	10°	70°
σ_4	25	51	51

3 RESULTS

Using the methodology described in the previous section, confusion matrices and Kappa coefficients were obtained for the performance of each method independently and for the fusion. This simulation was executed on a PC platform with a Intel i3 processor, 4GB of DDR3 RAM, and 1TB hard drive running Lubuntu v17.04. The total time taken to compute the result for the sample described before was 0.0031 seconds.

The FGauNB method individually presented Kappa of 39.55% and variance of 2.31×10^{-3} . This low agreement coefficient can be explained by the low amount of data assessed by this method, given that better results can be obtained when applied to more dimensions of data. The FBinNB method presented Kappa coefficient of 76.65% with variance 1.87×10^{-3} and the FExpNB method resulted on Kappa of 57.27% with variance 2.81×10^{-3} . Additionally, their confusion matrices are displayed on Table 9, 10 and 11, respectively.

As described before, after the results of each Fuzzy Naive Bayes method were obtained, weighted majority voting was used to perform the fusion of this methods. The resulting confusion matrix

Table 9: Confusion matrix for the FGauNB method.

	C1	C2	C3
C1	94	4	2
C2	82	15	3
C3	86	4	10

Table 10: Confusion matrix for the FBinNB method.

	C1	C2	C3
C1	80	11	9
C2	11	57	32
C3	0	7	93

Table 11: Confusion matrix for the FExpNB method.

	C1	C2	C3
C1	91	9	0
C2	54	40	6
C3	7	52	41

for the fusion is displayed on Table 12. The fusion resulted on Kappa coefficient of 79.98% with variance 1.85×10^{-3} , which means that the assessment was classified as substantial, according to Table 4.

Additionally, it is important to highlight that this assessment method never classified samples from class three, which is for unacceptable performance, as class one, which is for very good performance. This mistake would be critical since the main purpose of this system is to properly train physicians for real life cases.

Table 12: Resulting confusion matrix.

	C1	C2	C3
C1	80	12	8
C2	11	60	29
C3	0	9	91

4 CONCLUSION

Within this work, a new way of assessing data from virtual reality simulation for the gynecological exam was proposed. This was composed by the fusion determined methods assessing determined variables, depending on their statistical distribution. In order to evaluate if this is a

feasible assessment methodology, the assessment of this method was performed through its resulting confusion matrix and Kappa coefficient. Accordingly to the results obtained, this fusion method is feasible for assessment tasks, presenting a substantial agreement degree.

As future works, it would be interesting to analyze the behaviour of this fusion with more variables with different statistical distributions and methods, as well as different methods performing the fusion. Furthermore, we intend to implement the whole VR simulator for the gynecological examination.

ACKNOWLEDGEMENTS

This project is partially supported by grants 132170/2017-5 and 310561/2012-4 of the National Council for Scientific and Technological Development (CNPq).

RESUMO. xxxxxxxxxxxx

Palavras-chave: xxxxxxxxxxxx

REFERENCES

- [1] G. Burdea, G. Patounakis, V. Popescu & R.E. Weiss. Virtual Reality-based Training for the Diagnosis of Prostate Cancer. *IEEE Transactions on Biomedical Engineering*, **46**(10) (1999), 1253–1260.
- [2] G.C. Burdea & P. Coiffet. “Virtual Reality Technology”. John Wiley & Sons, 2nd edition (2003).
- [3] H. Carcio & R.M. Secor. “Advanced health assessment of women: Clinical skills and procedures”. Springer Publishing Company (2010), pp. 61–84.
- [4] J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychosocial Measurement*, **20**, 37-46 (1960).
- [5] A.D. dos Santos. “Simulação Médica Baseada em Realidade Virtual para Ensino e Treinamento em Ginecologia”. Master’s thesis, Universidade Federal da Paraíba (2010).
- [6] D. Dubois & H. Prade. Possibility theory: qualitative and quantitative aspects. In “Quantified representation of uncertainty and imprecision”. Springer (1998), pp. 169–226.
- [7] R.O. Duda, P.E. Hart & D.G. Stork. “Pattern classification”. John Wiley & Sons (2012).
- [8] M. Färber, E. Hoeborn, D. Dalek, F. Hummel, C. Gerloff, C.A. Bohn & H. Handels. Training and evaluation of lumbar punctures in a VR-environment using a 6DOF haptic device. *Studies in health technology and informatics*, **132** (2007), 112–114.
- [9] G.M. Foody. Status of land cover classification accuracy assessment. *Remote sensing of environment*, **80**(1) (2002), 185–201.
- [10] A. Gallagher, N. McClure, J. McGuigan, I. Crothers & J. Browning. Virtual reality training in laparoscopic surgery: a preliminary assessment of minimally invasive surgical trainer virtual reality (MIST VR). *Endoscopy*, **31**(04) (1999), 310–313.

- [11] J. Huang, S. Payandeh, P. Doris & I. Hajshirmohammadi. Fuzzy classification: towards evaluating performance on a surgical simulator. *Studies in health technology and informatics*, **111** (2005), 194–200.
- [12] INCA. INCA - Instituto Nacional de Câncer (2017). URL <https://mortalidade.inca.gov.br/MortalidadeWeb/>.
- [13] L.I. Kuncheva. “Combining pattern classifiers: methods and algorithms”. John Wiley & Sons (2004).
- [14] J.R. Landis & G.G. Koch. The measurement of observer agreement for categorical data. *biometrics*, (1977), 159–174.
- [15] L.D.S. Machado, R.M. De Moraes & M.K. Zuffo. Fuzzy rule-based evaluation for a haptic and stereo simulator for bone marrow harvest for transplant. In “5th Phantom Users Group Workshop Proceedings”. Citeseer (2000).
- [16] P.B. McBeth, A.J. Hodgson & M. Karim Qayumi. Quantitative Methodology of Evaluating Surgeon Performance in Laparoscopic Surgery. *Medicine Meets Virtual Reality 02/10: Digital Upgrades, Applying Moore’s Law to Health*, **85** (2002), 280.
- [17] R. Moraes & L. Machado. Another approach for fuzzy naive bayes applied on online training assessment in virtual reality simulators. In “Proceedings of Safety Health and Environmental World Congress” (2009), pp. 62–66.
- [18] R.M. Moraes & L.S. Machado. Assessment Systems for Training Based on Virtual Reality: A Comparison Study. *SBC Journal on 3D Interactive Systems*, **3**(1) (2012), 9–16.
- [19] R.M. Moraes & L.S. Machado. Online Assessment in Medical Simulators Based on Virtual Reality Using Fuzzy Gaussian Naive Bayes. *Journal of Multiple-Valued Logic & Soft Computing*, **18** (2012).
- [20] R.M. Moraes & L.S. Machado. Psychomotor skills assessment in medical training based on virtual reality using a Weighted Possibilistic approach. *Knowledge-Based Systems*, **70** (2014), 97–102.
- [21] R.M. Moraes & L.S. Machado. A Fuzzy Binomial Naive Bayes classifier for epidemiological data. In “Fuzzy Systems (FUZZ-IEEE), 2016 IEEE International Conference on”. IEEE (2016), pp. 745–750.
- [22] R.M. Moraes & L.S. Machado. A FUZZY EXPONENTIAL NAIVE BAYES CLASSIFIER. In “Uncertainty Modelling in Knowledge Engineering and Decision Making: Proceedings of the 12th International FLINS Conference (FLINS 2016)”, volume 10. World Scientific (2016), p. 207.
- [23] A. Papoulis & S.U. Pillai. “Probability, random variables, and stochastic processes”. Tata McGraw-Hill Education (2002).
- [24] J. Rosen, C. Richards, B. Hannaford & M. Sinanan. Hidden Markov models of minimally invasive surgery. *Studies in Health Technology and Informatics*, (2000), 279–285.
- [25] S.M. Ross. “Introduction to probability models”. Academic press (2014).
- [26] D. Ruta & B. Gabrys. An overview of classifier fusion methods. *Computing and Information systems*, **7**(1) (2000), 1–10.

- [27] E.A.d.M.G. Soares & R.M. Moraes. Assessment of Poisson Naive Bayes Classifier with Fuzzy Parameters Using Data from Different Statistical Distributions. (2016), 57–68.
- [28] H.P. Störr, Y. Xu & J. Choi. A compact fuzzy extension of the Naive Bayesian classification algorithm. In “Proceedings InTech/VJFuzzy” (2002), pp. 172–177.
- [29] H. Tahani & J. Keller. Information Fusion in Computer Vision Using Fuzzy Integral Operator. *IEEE Trans on Systems, Man and Cybernetics*, **20** (1990).
- [30] L.A. Zadeh. Fuzzy sets. *Information and control*, **8**(3) (1965), 338–353.
- [31] L.A. Zadeh. Probability measures of fuzzy events. *Journal of mathematical analysis and applications*, **23**(2) (1968), 421–427.