Taylor & Francis
Taylor & Francis Group

Journal of Interdisciplinary Mathematics

ISSN: 0972-0502 (Print) 2169-012X (Online) Journal homepage: https://www.tandfonline.com/loi/tjim20

Development of a new mathematical tool for early colorectal cancer diagnosis and its possible use in mass screening

Antonio Battista, Rosa Alessia Battista, Federica Battista, Luigi Cinquanta, Gerardo Iovane, Michele Corbisieri & Angelo Suozzo

To cite this article: Antonio Battista, Rosa Alessia Battista, Federica Battista, Luigi Cinquanta, Gerardo Iovane, Michele Corbisieri & Angelo Suozzo (2019) Development of a new mathematical tool for early colorectal cancer diagnosis and its possible use in mass screening, Journal of Interdisciplinary Mathematics, 22:6, 811-835, DOI: 10.1080/09720502.2019.1649834

To link to this article: https://doi.org/10.1080/09720502.2019.1649834

Published online: 27 Dec 2019.

Journal of Interdisciplinary Mathematics
ISSN: 0972-0502 (Print), ISSN: 2169-012X (Online)
Vol. 22 (2019), No. 6, pp. 811–835
DOI : 10.1080/09720502.2019.1649834

Taylor & Francis
Taylor & Francis Group

Development of a new mathematical tool for early colorectal cancer diagnosis and its possible use in mass screening

Antonio Battista
A. O. U. S. Giovanni di Dio e Ruggi d'Aragona
UOC Chir Urg
UOC Laboratorio
Analisi Salerno
Italy

Rosa Alessia Battista
Federica Battista
University Vita Salute S. Raffaele
IRCCS Ospedale San Raffaele
Milano
Italy

Luigi Cinquanta
A. O. U. S. Giovanni di Dio e Ruggi d'Aragona
UOC Chir Urg
UOC Laboratorio
Analisi Salerno
Italy

Gerardo Iovane *
Michele Corbisieri
Angelo Suozzo
Dipartimento di Informatica
University of Salerno
Salerno
Fisciano 84084
Italy

*E-mail: giovane@unisa.it

Abstract

The aim of the study is to model a new mathematical tool effective in the diagnosis of colorectal cancer and useful as a mass screening test. This work enrolled 345 subjects from January 2011 to January 2013, in 4 Italian Surgical Departments: 97 (27,4%) out of the total were healthy controls (class 0) whereas 248 (72,6%) patients were affected by colorectal cancer. The cancer patients were divided into four classes, according to TNM staging classification: 74 (20,9%) patients stage I (class 1); 61 (17,2) patients stage II (class 2); 76 (21,5%) patients stage III (class 3); 37 (10,5%) stage IV (class 4).

Blood samples were collected after 24h from hospital admission and 17 biochemical parameters (CEA, ceruloplasmin, haptoglobin, transferrin, TPA, CA 19.9, CA 72.4, PCR, Ca 50, C4 Complement, CA 125, Alfa-1- antitrypsin, alpha-2-Macroglobulin, Ferritin, RBP, Alpha-1-Acid Glycoprotein, Complement C3) were statistically analysed together with the clinical and pathological disease staging (pre and postoperative evaluation, respectively). Evaluation and comparison were made between two groups: healthy controls and the total of affected patients.

Using the collected data, it was developed a mathematical model (artificial neural network ANN) allowing the distribution of colorectal cancer and controls (apparently in good health) patients in the two groups. This led to the definition of an index (B- index) that, simply analysing the combined values of the 17 biochemical parameters, decides on the either healthy or disease status of the patient. The use of the B-index, with a neural network based on real values, allows colorectal cancer diagnosis with 80.078% accuracy, probability of false positive (FP) = 0.333, probability of false negatives (FN) = 0.102, a 0.898 sensitivity and 0.667 specificity. However, by using B-index with a neural network based on the implementation of extended reality values our analysis reaches an accuracy of 91.11% in colorectal cancer diagnosis, with probability of false positive (FP) = 0.472, probability of false negatives (FN) = 0.03%, a 0.9997 sensitivity and 0.7642 specificity. The results suggest a promising role for B-index in colorectal cancer mass screening with an easily available, low cost and non-invasive test.

Subject Classification: *Dynamical Systems, Applications of Mathematics*

Keywords: *B-index, Colorectal cancer diagnosis, Artificial intelligence, Simulation*

1. Introduction

Colorectal cancer is the third most common cancer in men and the second in women worldwide, with an incidence of nearly 700.000 cases per year. Mortality accounts for 8% of all cancer related deaths whereas 5-year survival rate is about 65% over all patients affected by this type of tumour. However, a decrease in incidence and mortality rates was observed over the past decades, along with an increase of early diagnosis compared to late ones. This trend seems to be highly related to the introduction of useful tools for mass screening. Currently, recommended screening tests for at risk population include stool-based tests and endoscopic ones. However,

they were influenced by low sensitivity on one hand, while on the other hand they implicate quite invasive procedures. In this perspective research should strive to delineate new diagnostic approaches, aiming earlier diagnosis and better cost-effectiveness procedures as mass screening tests.

The aim of this work goes along these lines and it is focused on designing a mathematical model able to distinguish the patients apparently healthy (class 0) from colorectal cancer patients (class 1: TNM stage I-IV) by using serum level of 17 relevant biomarkers (CEA, ceruloplasmin, haptoglobin, transferrin, TPA, CA 19.9, CA 72.4, PCR, Ca 50, C4 Complement, CA 125, Alfa-1- antitrypsin, alpha-2-Macroglobulin, Ferritin, RBP, Alpha-1-Acid Glycoprotein, Complement C3). This new clinical tool, named B-index, would lead to early diagnosis of colorectal cancer with a promising role in mass screening.

From January 2011 to January 2013, 345 patients were enrolled in this study: 97 healthy controls and 248 patients affected by colorectal cancer. Patients suffering from colorectal neoplastic disease were enrolled at the following four Surgical Departments

- Giovanni di Dio e Ruggi dAragonaHospital, Salerno
- Advanced Specialized â€œMonaldiHospital, Naples
- anto SpiritoHospital, Bra (CN)
- MatteoHospital, Pavia

Preoperative blood sampling was performed by venipuncture after 24 hours from hospital admission in conditions of reduced psychological stress.

The developed tool is an artificial neural network trained with data coming from 89 patients whereas the remaining 256 were used for testing the accuracy of the obtained network. The network has two outputs: healthy state and disease state (Authors intention is the critical evaluation of the possible further implementation with a 5 outputs B-index (0: apparent good health; 1: TNM stage I; 2: TNM stage II; 3: TNM stage III; 4: TNM stage IV)). Healthy controls were selected without limitations of sex and age (between 45 and 80 years). The affected patients were selected without limitations of sex and were aged 45-87 years. They underwent preoperative staging including RCS, biopsy, histological analysis, total-body CT with contrast medium. Postoperative staging was defined through surgical histology, and TNM final definition. Thus, they were divided according to their disease stage: 74 patients stage I; 61 patients stage II; 76 patients stage III; 37 patients stage IV based on TNM Blood samples were collected in all

patients with dosage of the following serum parameters: CEA [19, 16, 30, 21], ceruloplasmin [22], haptoglobin [33, 25], transferrin [32, 27, 9], TPA [28], CA 19.9 [1, 24, 13], CA 72.4 [15, 34], PCR [10, 20], CA 50 [1, 26], C4 Complement [12, 17], CA 125 [29, 3, 4], Alpha-1-antitrypsin [33], alpha-2 -Macroglobulina [7, 17], Ferritin [27, 14, 11, 23], RBP [18, 5], Alpha- 1-Acid Glycoprotein [7], Complement C3 [17, 2].

Differences between the doses of preoperative tumour markers and acute phase proteins were studied with evaluation of mean, STDEV and t-Student in the two groups (patients in apparent good health vs patients with neoplastic disease stage I-IV) [19, 8, 31, 13, 7]. Development of the neural network, with simultaneous evaluation of all the serum proteins and antigens studied [28, 6, 31, 34], was made with two outputs (state of absence of disease (controls) - state of the disease (stage I - IV TNM)).

2. The problem

Let $\mathcal{O}$ be a living organism in a biological system whose complexity is not known a priori and let $\mathcal{S}$ be a biological district, or rather a group of districts or biological subsystems of $\mathcal{O}$ such that $\mathcal{O}$ includes or coincides with $\mathcal{S}$. Given that the organism $\mathcal{O}$ is the human body, the behaviour of the biological system $\mathcal{S}$ is defined by the states that it can assume according to TNM standard staging of colorectal cancer. According to the assumptions made, the variables of $\mathcal{S}$ which identify the internal state of the system are known in terms of quantity or concentrations of proteins; they are also known the corresponding output values of the biological parameters that allow the staging. The very definition of organism (set of parts structured and interconnected) establishes that the candidate system has a not chaotic random nature and this allows to treat the system as described by theory of physical systems.

The system $\mathcal{S}$ is, obviously, a dynamic system which evolves over time due to external stimulus. It will be proved that $\mathcal{S}$ can be represented as a finite state system.

A finite state system is a discrete and stationary system whose inputs, outputs and states can take only a finite number of values and configurations.

The system $\mathcal{S}$, shown in the figure 1, can be described by a set of internal parameters (e.g. weight, volume, density, temperature) that completely represent the features of the system or, in other words, the state. Although to describe many physical objects or natural phenomena a representation of infinite dimension is required, in practice the prevalence of methods for the analysis and the synthesis of fixed size systems paves the

Figure 1

Representation of the system

way to the employment of approximations (e.g. time discretization) which allow to replace an infinite representation with a finite representation. This approximation can be also applied to the system $\mathcal{S}$: indeed, the states, the inputs and the outputs can be observed and measured only in a fixed time slots and, as well, a fixed representation of values may be adopted in order to have a system with a finite cardinality.

The system $\mathcal{S}$ reacts to the stimuli from the surrounding environment by generating an output whose characteristics depend, generally, on the input value at that time, on the inner state of the system at that time and on its free evolution which is the sequence of intermediate states between the starting state when the stimulus appears and the final state when the output is sensed.

Let $\mathcal{U}$ and $\mathcal{Y}$ be, respectively, the set of symbols of input and output and let $\mathcal{X}$ be the set of symbols of the state; the implicit representation of the system $\mathcal{S}$ is given by:

$$x(k+1) = f(k, x(k), u(k)) \quad (1)$$

$$y(k) = \eta(k, x(k), u(k)) \quad (2)$$

where $f : \mathbb{N} \times \mathcal{X} \times \mathcal{U} \to \mathcal{X}$ is the generation or next state function and $\mu : \mathbb{N} \times \mathcal{X} \times \mathcal{U} \to \mathcal{Y}$ is the output function.

According to the biological assumptions made, the output of the system described in this study is independent of the input values; therefore, the system is defined proper and the equations (1) and (2) become:

$$x(k+1) = f(k, x(k), u(k)) \quad (3)$$

$$y(k) = \eta(k, x(k)) \quad (4)$$

If the laws that represent the intrinsic functioning of the system are constant over time, the direct dependence of time disappear from the characteristic equations:

$$x(k+1) = f(x(k), u(k)) \quad (5)$$

$$y(k) = \eta(x(k)) \quad (6)$$

Conversely, on the assumption that the cancer organism undergoes a significant alteration of microscopic processes (e.g. molecular synthesis)

whose scale also impacts on macroscopic processes, a new distinct system for each way in which the organism works should be defined. However, also in this scenario the study of the biological system $\mathcal{S}$ can be reduced to the study of n steady-state systems, each of them defined by a couple of equations such as (5) e (6), in which the functions f and μ take specific expressions due to a pathological alteration of cognitive processes.

In the light of the above, the system $\mathcal{S}$ is a discrete stationary system or rather a finite state system which could be reduced to n finite state systems where n is the number of distinct terms that S can assume. In many cases, the sets $\mathcal{U} = \{u_1, u_2, ..., u_m\}$, $\mathcal{X} = \{x_1, x_2, ..., x_p\}$ and $\mathcal{Y} = \{y_1, y_2, ..., y_r\}$ are algebraically unstructured in whole or in part; therefore, the functions f and η cannot be analytically assigned. Indeed, useful analytical representations for the study of the system can be achieved when a mathematical theory can be defined on the previous sets.

In the following, the system will be represented as a black box due to the difficulty of decompose the system in elementary parts whose mathematical relations (5) e (6) can be known; therefore, only properties, behaviours and trends of the system can be defined according to physical-pathologically knowledge.

The main purpose of this study is addressed to solve the following problem:

Identify the wellness or the cancer (according to TNM standard) of a generic subject (a specific instance of the organism $\mathcal{O}$) starting from the measurements of some internal parameters that are characteristic of the system $\mathcal{S}$ and belonging to the set $\mathcal{X}$ of its inner state variables, given that $\mathcal{S} \subseteq \mathcal{O}$.

Each element of the set $\mathcal{X}$ is a vector of n components and therefore $\mathcal{X}$ is a set of vectors:

$$\mathcal{X} = \{X(x_1, x_2, \ldots, x_n) \mid x_1, x_2, \ldots, x_n \in \mathbb{R}\} \tag{7}$$

where $x_1, x_2, ..., x_n$ are the values of the proteins measured in the single subject.

Following the TNM standard, let $y_1, y_2, ..., y_l$ be the l parameters observed and measured for the clinical or pathological classification of the single patient according respectively to cTNM and pTNM in relation to T - Primary Tumour, N - Regional Lynph Nodes and M - Distant Metastasis categories. Let $\widetilde{\mathcal{Y}}$ be the set containing all possible l-ple values that parameters $y_1, y_2, ..., y_l$ can assume in subjects classifiable in TNM stages I, II, III and IV:

$$\begin{aligned} \widetilde{\mathcal{Y}} &= \{Y(y_1, y_2, \ldots, y_l) \,|\, y_1, y_2, \ldots, y_l \in \mathbb{R} \ni (y_1, y_2, \ldots, y_l) \\ \Rightarrow \quad TNM\,stage &= C\} \end{aligned} \tag{8}$$

Similarly, $\mathcal{Y}_0$ will denote the set whose elements are all the possible l-ple of values that the parameters y_1, y_2, ..., y_l can assume in subjects not suffering from colorectal cancer and, therefore, are not associated to any of TNM stages:

$$\begin{aligned} \mathcal{Y}_0 &= \{Y(y_1, y_2, \ldots, y_l) \,|\, y_1, y_2, \ldots, y_l \in \mathbb{R} \ni (y_1, y_2, \ldots, y_l) \\ \Rightarrow \quad TNM\ stage &= \varnothing\} \end{aligned} \tag{9}$$

It should be noted that $\mathcal{Y}_0$ is the complement of $\widetilde{\mathcal{Y}}$, and therefore

$$\mathcal{Y} = \widetilde{\mathcal{Y}} \cup \mathcal{Y}_0 = \widetilde{\mathcal{Y}} \cup C(\widetilde{\mathcal{Y}}) \tag{10}$$

is, according to set theory postulates, the universe of values that both healthy and colorectal cancer patients can assume according to TNM standard.

Let $\psi : Y \in \mathbb{R}^l \to \mathcal{Y}$ be an application whose explicit form is:

$$\mathcal{Y} = \psi(Y) \tag{11}$$

Hence, the transformation θ is defined as:

$$\theta = \begin{cases} \varnothing, & \text{if} \quad \mathcal{Y} \equiv \mathcal{Y}_0 \\ C, & \text{if} \quad \mathcal{Y} \equiv \widetilde{\mathcal{Y}} \end{cases} \tag{12}$$

or, in the explicit form

$$\Psi = \theta(\mathcal{Y}) \tag{13}$$

where $\Psi \in \{\varnothing, C\}$.

By replacing the equation (11) in (13) the following relation is obtained:

$$\begin{aligned} \Psi(k) &= \theta(\psi(Y)) = \theta(\psi(\eta(X(k)))) = \gamma(X(k)), \\ \forall k &= 0, 1, 2, \ldots, r < \infty \end{aligned} \tag{14}$$

The system $\mathcal{S}$ is so characterized by the equations:

$$X(k+1) = f(X(k), U(k)) \tag{15}$$

$$\Psi(k) = \eta(X(k)) \tag{16}$$

The physical model that represents the subject of study was just defined on the basis of problem's parameters: indeed, the state and the output of $\mathcal{S}$ are known in a steady state and in some fixed times k; this means that the values of measured proteins (quantities or concentrations)

and the contextual colorectal tumour stage are known for every patient belonging to selected sample. Starting from these sampled data this work tries to establish the general behaviour of $\mathcal{S}$ at any instant k.

In this study the definition of the time evolution of the system is not a central concern: indeed, the aim of the work is not to synthesize a model that describes the evolution of a healthy subject into a patient with a colon cancer or the progression or regression of tumour in a sick subject. Therefore, the determination of the next state function *f* will be neglected in this study.

According to arguments and biological hypothesis presented, the problem of identification of a biological system $\mathcal{S}$ represented by (15) and (16) can be reduced to the synthesis of the function γ starting from the knowledge of the couples (X, Ψ), made up from, respectively, the vector whose components are the parameters identification of the internal state of $\mathcal{S}$ (the n values of the dosed proteins) and from the symbol defined in bi-univocal correspondence with the sets of vectors whose components are the l parameters taken into account to perform the TNM standard staging.

The impossibility of analytically represent the function γ motivates the employment of an artificial neural network to synthesize the function.

3. The solution

3.1 *Modelling via artificial neural network (ANN)*

A pre-processing of the input vector X was performed: the values of each proteins were centred on the zero because the presence of both positive and negative values improves the strengthening and weakening mechanisms of the synaptic values; therefore, the variability intervals were scaled for the variance of each distribution, obtaining for them comparable extensions.

The symbols belonging to the set of outputs were so coded:

$$\varnothing = (1,0); C = (0,1)$$

One can therefore conclude that the output has values in:

$$\Psi \in (1,0),(0,1)$$

A supervised feed-forward ANN will be used; indeed, a typical employment for this kind of networks is addressed to solve problems of pattern recognition and classification. The solution of the problem presented above is to achieve a network able to classify each input to the corresponding output category.

3.2 *Architecture and training of ANN*

Previously it has been proved that the system subject of study is an automaton or, in other words, a finite steady-state system according to Systems Theory definitions. A system is defined stationary if the laws that govern its behaviour are time-invariant. This statement, obviously, cannot be applied to an organism which develops or recovers from a tumour: in both cases, the alterations that occur involve the intrinsic microscopic and macroscopic nature (genetic behaviour, etc.). One can then assert that a mathematical model that is able to describe the behaviour both in the presence and absence of the cancer cannot be applied to the system $\mathcal{S}$. Or rather, the behaviour of the system cannot be described by using only one system: a new automaton can be synthesized to proper describe each evolution of a subject with its own characteristics.

The artificial neural network was realized as the parallel of two distinct networks (equal to the number of automata identified according to bio-physical criteria) each of them equipped with hidden layers that share the inputs from the 13 input units (totally connected to the hidden neurons) and return the output to the 2 output neurons, each of them specialized to encode a component of the vector Ψ.

The training phase learns to the ANN the association that exists (if exists) between the input and the output vector. The task of the supervisor was to provide to network the events that exhaustively represent in terms of number and type the nature of the problem subject of the study. The patterns chosen for the training were the result of a careful selection made by the supervisor; the ANN training set was made up by 89 subjects among the 345 individuals of the entire sample. Each subject was identified by a couple made up by a vector of 13 proteins and by the relative TNM staging class. In order to prevent over-fitting and preserve the generalization ability of the network, 28 additional subjects whose characteristics are similar to those of the training set were chosen and equally divided into validation and testing sets. The latter were used in the training phase of the ANN to identify the stop criterion (early stop) of the algorithm of back-propagation.

3.3 *Presentation of the results*

Once that training dataset was chosen, many tests were performed to identify the neural network that performs the best the diagnostic test. Specifically, many tests are made on networks with different inner layers: the number of nodes was progressively varied by increasing or decreasing

small amounts to evaluate the behaviour exhibit by the network. The activation functions were varied, as well; moreover, the values of the weights at the beginning of the training were set using a function that generate pseudo-random numbers according to a uniform distribution. The identified RNA used to simulate the output function γ of the system $\mathcal{S}$ has provided the following results.

Le T^+ and T^- be the events "positive test result" ($\Psi = C$) and "negative test result" ($\Psi - \emptyset$) respectively. According to the probability theory, $P(A \mid B)$ denotes the probability that the event A occurs given that the event B has already occurred. Moreover, let FN be the number of false negatives, FP the number of false positives and NC the number of subjects for which the RNA was unable to provide a classification. By extension, let $\emptyset$ be the event "the subject belongs to the group of controls " and let C be the event "the subject is affected by colon cancer",

System = $\mathcal{S}$

Output function = γ

Sample size = 345

Learning set size = 89

Validation set size = 14

Test set size = 14

Number of performed simulations = 256

Number of successful simulations = 205

Number of failed simulations = 51

of which

Number of NC = 11

Number of FN = 20

Number of FP = 20

The diagnostic test has the following statistical characteristics:

Performance = 80.078%

Sensitivity = $P(T^+ \mid C) = 0, 898$

Specificity = $P(T^- \mid \emptyset) = 0.667$

P (FP) = $P(T^+ \mid \emptyset) = 0.333$

P (FN) = $P(T^- \mid C) = 0.102$

At this point, a cognitive learning was necessary for a further implementation.

4. The enforcement of statistics via extended reality simulation

4.1 *Introduction*

Typically, when the statistic is not enough there are several ways to perform more detailed analysis on larger samples by employing two main types of simulations: stochastic and extended reality simulations. The first implies the generation of a new sample completely disclosed from the set of original data. In this case, the generation is performed according to the distribution function chosen by the user. This approach is particularly useful to verify if the configuration of the observed data has a specific trend, characteristic features, etc. compared to a pure probabilistic sample such as that generated by the simulation.

The extended reality simulations are much more conditioned because they require that the original data set plays a key role in the simulated set. Typical examples are the refinement of probability distributions in which the simulated sample must have the same probability distribution, or the population of a class distribution does not change; another employment for this simulation is addressed to generate random values confined in the limit of variation bands corresponding to those observed and so on.

The purpose of these simulations is therefore to generate a larger sample than the original one, in order to study the differentiation properties that are not visible in low statistics but having the same general properties of the original sample which have less data.

The expected behaviour of the extended reality simulations - from a probabilistic point of view - can be of three types:

(1) Weakening of the observed signal and of the signal to noise ratio;
(2) Confirmation of the observed signal and of the signal to noise ratio;
(3) Strengthening of the observed signal and of the signal to noise ratio.

The occurrence of the condition 1 is indicative of a low significance of the obtained signal and of a low representativeness of the used sample compared to the population or statistical universe to which the sample belongs.

The occurrence of the condition 2 is indicative of a fully representativeness of the sample compared to the population.

The occurrence of the condition 3 is indicative of a good representativeness of the original sample; the employment of the extended reality simulation allows to identify new characterizations of

the population, to better differentiate the phenomenon and to increase the ability to distinguish the noise signal when the signal to noise ratio improves during the execution of the simulation in relation to the various optimizations considered in the simulation of extended reality.

4.2 *The purpose of the extended reality and of chosen constraints*

In this study an extended reality simulation was made to prove or disprove the interesting result obtained with original dataset. The established constraints were the following:

(1) The generated data will be generated according to the TNM classes: this is a basic requirement to avoid the creation of hybridizations between different stages of the cancer;

(2) Fuzzification of diagnostic parameters. The frequency distribution and the variation of the parameters were analysed among the entire range of variation, where the latter was divided into five classes:

- 1 = low;
- 2 = below average;
- 3 = average;
- 4 = above average;
- 5 = high

(3) The value of each parameter for each diagnostic profile (that is for each subject considered in the original set) was remapped. The advantage of this fuzzification process consists in reduce the computational cost, without affect the ability to have the same high statistical representation in the set of final data. Then, the fuzzified string of subject is build starting from the 13 parameters that, this time, are not real numbers but integer values in the fuzzy set as above explicate.

(4) Conditioned stochastic generation. In this phase, given the TNM, all possible combinations of the 13 parameters taken into account are created. The number of cases that can be generated varies depending on the type of generation to be undertaken. Specifically, the achievable significant simulations may generate a new sample with a number of cases that varies in the range of [13^5, 1, 18^*10^{33}], or between 371.293 cases (ie 13^5 or by randomly changing within classes identified in the previous fuzzification process) and 1, 18^*10^3 cases (corresponding to 350^{13}, the most general possible case). This interval clearly sets the

lower bound that allow to perform a simulation of extended reality in which the sample is fully representative of the universe whereas the computational cost is limited although the upper limit cost is not feasible for the current computing power even if with a view to Big Data. However, the world population is in the order of 6, $5^{*}10^{9}$ and therefore far less than the upper limit of 10^{33} cases.

(5) Data Generation for realistic simulation. Once 400,000 cases ware generated, the following steps were followed. Each datum, before this phase is a n-uple of 13 elements with values range from 1 to 5. With this step the integer values in the range [1,5] are converted into real values using a random inter-class generation.

Example :

- **Step 1 :** A parameter varies in the range [0,99] belonging to the set of real numbers, given the TNM.
- **Step 2 :** The [0.99] range is divided into the following five classes
  (1) low = [0.20 [;
  (2) below average = [20,40 [;
  (3) average = [40,60 [;
  (4) above average = [60,80 [;
  (5) high = [80,99]
- **Step 3 :** In this way the string of 13 parameters (real numbers) can be transformed into a string of 13 integer values that can take a value between 1 and 5.
- **Step 4 :** In order to lock the statistics of ** possible cases to 400,000 cases for each class, a hierarchical sorting is introduced in which not all the recombinations of the parameters are allowed since the relative generation step is stopped when the limit of 400,000 cases is reached. Therefore, the dominance hierarchy of parameters that may change first is the following, from most important to least important parameter:
  (1) CERULOPLAS
  (2) CEA
  (3) CA72-4
  (4) CA19-9
  (5) CA50

(6) CA125
(7) TPA
(8) APTOGL
(9) A1ANTITRIPS
(10) A 2 MACROGL
(11) PCR
(12) GLICOPROT AC
(13) RBP
(14) TRANSF
(15) FERRITIN
(16) C3
(17) C4

Specifically, the parameters from 14th to 17th were not involved in recombination, in order to reduce the cardinality of the statistic but they were remapped exactly with the class value of the original data set (this clearly without loss of generality). As an example, one can think to have a sample that varies between z.

- **Step 5** - In the simulated realistic generation, each parameter is mapped in its original belonging interval. For example, in the case that the fixed parameter in the previous step had a score of 4 as fuzzy parameter it may assume a value in the range [60, 80[. In this phase a random number in that range will be generated. Let 68,75 be, for example, the obtained value.

That generation will be performed for each parameter and for each recombination generated in the previous phase. This will lead to transform the 400,000 fuzzified cases obtained in the previous stage to 400,000 cases with realistic values for each given TNM.

4.3 *Software Module for the realistic simulation*

The developed software has achieved the goal of process data relating to the clinical study of 346 patients. The data on the clinical study are stored on a "csv" file, which is constituted by:

- 346 lines, where each line is a patient;
- 21 columns, where each column is a clinical parameter.

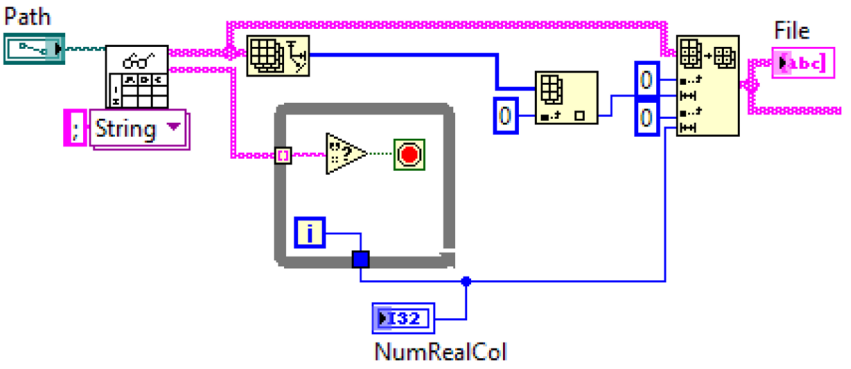

Figure 2

CSV File Reading

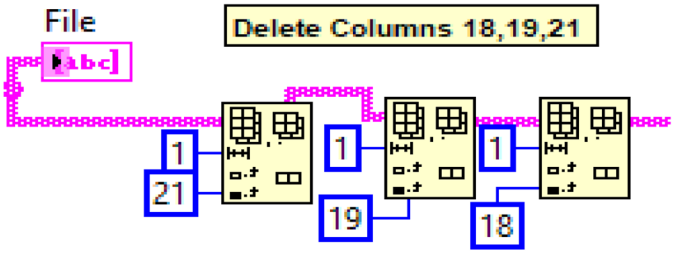

Figure 3

Columns to neglect

Specifically, the last 4 columns are not of significant interest at this stage of the study because concerning the TNM (in this study the classes of TNM, the patient's age, sex and an internal classification parameter linked to neural network which has been discussed are not mixed).

The objective of the software is to develop a realistic simulator with the aim to build all possible combinations, where each row will always have 21 elements.

In Figure 2, the input file is read and in Figure 3 the columns corresponding to the parameters Gender, Age and Ann Group are deleted because they are considered not relevant in this study.

In order to reduce the computational cost of the software execution a patient classification function has been realized according to the TNM value of the parameter. This parameter can assume values equals to: 0, 1, 2, 3 and 4.

The function shown in Figure 4 performs the TNM classification by taking in input the matrix containing the data and returns in output the following data structures:

- In TNM0 data structure only patients with a TNM value of 0 are stored ;
- In TNM1 data structure only patients with a TNM value of 1 are stored ;
- In TNM2 data structure only patients with a TNM value of 2 are stored ;
- In TNM3 data structure only patients with a TNM value of 3 are stored ;
- In TNM4 data structure only patients with a TNM value of 4 are stored ;
- Finally, in SplitTNM data structure all patients are stored according to TNM.

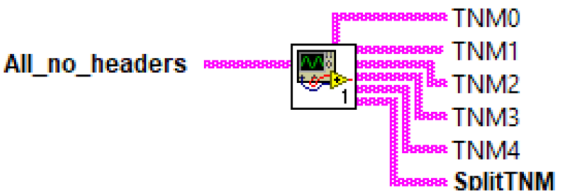

Figure 4

TNM division

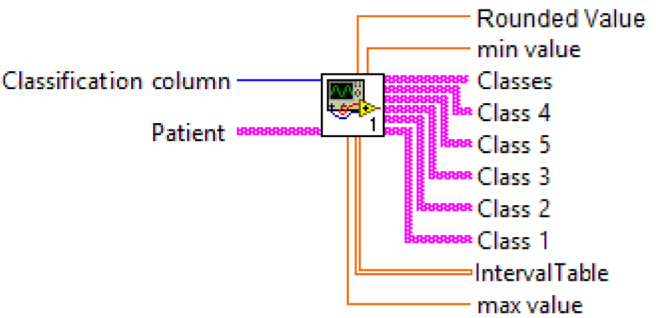

Figure 5

Patients classification

Figure 6

Classifier Split TNM

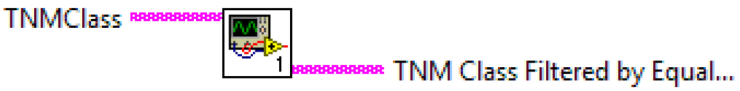

Figure 7

TNM filter

The TNM division division is used as input for the realization of a classifier (Figure), which has the aim of:

- Classifying patients into five distinct classes;
- Generate a table containing the values of the ranges for each class.

The idea behind the classifier achieved is the following: for each parameter (column) is calculated the following value: $(\text{Max} - \text{Min})/5$.

Considering the obtained value (Rounded Value) and the respective maximum and minimum values (max value and min value) five classes of patients will be identified:

The other outputs shown in Figure 4 concerning data structures TNM0, TNM1, TNM2, TNM3 e TNM4 show the following TNM division:

- 97 patients with TNM equal to 0;
- 74 patients with TNM equal to 1;
- 61 patients with TNM equal to 2;
- 76 patients with TNM equal to 3;
- 37 patients with TNM equal to 4.

For each of data structure a further function is applied which carries out the classification, as is shown in Figure 6.

With the aim to further reduce the computational cost of the software a filter function (filter on TNM), shown in Figure 7, has been realized.

The realized function has the role to ignore the patients with the same class (the equal records are not taken into account). The result obtained by this function has significantly reduced the number of patients for each TNM class, thus reducing the computational cost of the application.

After that the patients were classified as described before, the part relating to the simulation of all possible combinations was realized. To speed up the execution of the software, the user can choose if perform the combined calculation (Figure 8):

- On all parameters;
- On 16 parameters;
- On 15 parameters;
- On 14 parameters;
- On 13 parameters;
- On 12 parameters;
- On 11 parameters;
- On 10 parameters;
- On 9 parameters;
- On 8 parameters;
- On 7 parameters;
- On 6 parameters;
- On 5 parameters.

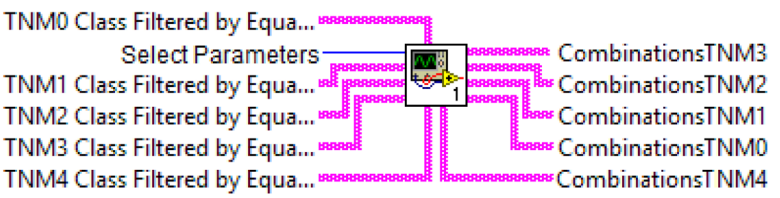

Figure 8

Combinations generation

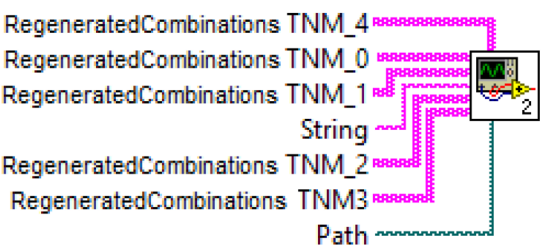

Figure 9

CSV file writing

Figure 10

Classes of patients

Figure 11

Classes per TNM

Figure 12

Combinations output per TNM

Finally, after obtaining all the possible combinations of TNM, the results are written on a ".csv" file, as shown in Figure 9.

This function takes in input:

- The combinations obtained by the software;
- A string containing the prefix to concatenate to files to distinguish the different versions;
- The relative path to the directory containing the ".csv" file.

The example shown below was performed on a test file containing 10 records (10 patients). In Figure 10, the data structure containing all classes is shown.

In Figure 11, the resulting classes after performing the division by TNM are shown. The data structure containing the classes with TNM equal to 3 is empty because the input file given to the software has no patients with a TNM equal to 3. Finally, the Figure 12 shows the output of the combinations, classified by TNM, obtained combining only five parameters.

4.4 *Results from extended reality simulation*

The results of this study presented in this section fall into the case 3 presented in section 4.1, that is the strengthening of the observed signal and the signal to noise ratio.

Specifically, the results table reported below must be compared with that obtained in the case of the original data set (346 subjects), from which the signal enhancement is clearly proofed due to differentiation that is realized when 2 million of cases are considered.

- Success factor = 91.11%
- False Positives = 0.472%
- False Negatives = 0.03%
- Sensitivity = 0.9997
- Specificity = 0.7642
- Pos.TEST PRED = 0.01468
- Neg.TEST PRED = 1

5. Results and conclusions

From 2011 to 2013, 345 patients (97 healthy and 248 affected by colorectal cancer) without gender differentiation and aged between 45 and 87 were pre-operative dosed, 24 hours after hospitalisation in stable and normalized conditions with regard to stress, CEA, Ceruloplasmin, Haptoglobin, Transferrin, TPA, CA 19.9, CA 72.4, PCR, Ca 50, C4

Table 1

Protein	Control vs k stages I-IV (p)	Protein	Control vs k stages I-IV (p)
CEA	<< 0.0001	Complement C3	< 0.001
Ceruloplasmin	<< 0.0001	C4 - Complement	< 0.05
Haptoglobin	<< 0.0001	CA 125	< 0.05
Transferrin	< 0.05	Alpha-1-antitrypsin	< 0.001
TPA	< 0.05	Alpha-2-macroglobulin	<< 0.0001
CA 19.9	<< 0.0001	Ferritin	< 0.001
CA 72.4	<< 0.0001	RBP	< 0.001
PCR	<< 0.0001	Alpha-1-acid glycoprotein	< 0.0001
CA 50	<< 0.0001		

Table 2

Protein	R^2	Protein	R^2
CEA	0.25	Complement C3	0.09
Ceruloplasmin	0.42	C4 - Complement	0.06
Haptoglobin	0.14	CA 125	0.06
Transferrin	0.06	Alpha-1-antitrypsin	0.09
TPA	0.05	Alpha-2-macroglobulin	0.15
CA 19.9	0.13	Ferritin	0.04
CA 72.4	0.19	RBP	0.10
PCR	0.12	Alpha-1-acid glycoprotein	0.16
CA 50	0.13		

Complement, CA 125, Alpha-1-antitrypsin, Alpha-2-macroglobulin, Ferritin, RBP, Alpha-1-acid glycoprotein, C3 Complement.

The differences between pre-operative cancer markers doses and acute phase proteins were studied with an average evaluation, standard deviation and t Student in two groups (healthy patients vs tumour I-IV stage) with significant statistical difference (tab. 1).

Another evaluation done was relative to pre-operative doses trend in apparently healthy patients vs cancer patients (TNM I-IV stage) with simple linear correlation method between doses and cancer stage (tab. 2).

A successive evaluation was made by the ANN with the simultaneous evaluation of all serum proteins and antigen studied. To realize the ANN

the data of 89 patients were used whereas the remaining 256 were used to check the performances of the network.

The network with 2 output (B-index: state of health and state of sickness) has realized a performance of 80.078%, probability of false positive P(FP) = 0.333, probability of false negative P(FN) = 0.102 with a sensibility of 0.898 and a specificity of 0.667. The assimilation of false positive to the good usage of the network could lead to a performance of 87.5%, by considering that a false positive always allows further analysis to establish the correctness of pre-operative analysis.

However, the condition that makes this study still more interesting is the performance achieved by the ANN with the employment of "extended reality simulation", previously widely discussed. In fact, the result of the network with two outputs were:

- Success factor= 91,11%
- False Positives = 4.72%
- False Negatives = 0.03%
- Sensitivity = 0.9997
- Specificity = 0.7642
- positive predictive value = 0.01468
- negative predictive value = 1

These results make very useful the employment of the B-index method in mass screening which is preparatory for further radiological or instrumental analysis addressed to early diagnosis of colorectal cancer.

References

[1] JA Alvarez et al. "Sensitivity of monoclonal antibodies to carcinoembryonic antigen, tissue polypeptide antigen, alpha-fetoprotein, carbohydrate antigen 50, and carbohydrate antigen 19-9 in the diagnosis of colorectal adenocarcinoma". In: *Diseases of the colon & rectum* 38.5 (1995), pp. 535–542.

[2] A Archimandritis et al. "Serum Protein Markers (Hp, GC, C3) in Patients with Colon Cancer". In: *Human heredity* 43.1 (1993), pp. 66-68.

[3] RC Bast Jr et al. "CA 125: the past and the future". In: *The International journal of biological markers* 13.4 (1998), pp. 179–187.

[4] J Bukowski et al. "CA 19-9 and CA 125 antigens in the sera of patients with cancer of the large intestine in relation to its clinical progress". In: *Wiadomosci lekarskie* (*Warsaw, Poland*: 1960) 42.1 (1989), pp. 30–34.

[5] P Charpiot et al. "Vitamin A, vitamin E, retinol binding protein (RBP), and prealbumin in digestive cancers." In: *International journal for vitamin and nutrition research. Internationale Zeitschrift fur Vitamin- und Ernahrungsforschung. Journal international de vitaminologie et de nutrition* 59.4 (1989), pp. 323–328.

[6] Chien-Chih Chen et al. "Is it reasonable to add preoperative serum level of CEA and CA 19-9 to staging for colorectal cancer?" In: *Journal of surgical research* 124.2 (2005), pp. 169–174.

[7] P Durdey, NS Williams, and DA Brown. "Serum carcinoembryonic antigen and acute phase reactant proteins in the pre-operative detection of fixation of colorectal tumours". In: *British journal of surgery* 71.11 (1984), pp. 881–884.

[8] G Gebauer and W Müller-Ruchholtz. "Tumor marker concentrations in normal and malignant tissues of colorectal cancer patients and their prognostic relevance." In: *Anticancer research* 17.4A (1997), pp. 2731–2734.

[9] Lisa J Herrinton et al. "Transferrin saturation and risk of cancer". In: *American journal of epidemiology* 142.7 (1995), pp. 692–698.

[10] Christian Kersten et al. "Increased C-reactive protein implies a poorer stage specific prognosis in colon cancer". In: *Acta oncologica* 52.8 (2013), pp. 1691–1698.

[11] Teruyuki Kishida et al. "Clinical significance of serum iron and ferritin in patients with colorectal cancer". In: *Journal of gastroenterology* 29.1 (1994), pp. 19–23.

[12] Ken Konishi et al. "Expression of C4. 4A at the invasive front is a novel prognos tic marker for disease recurrence of colorectal cancer". In: *Cancer science* 101.10 (2010), pp. 2269–2277.

[13] P Kuusela et al. "Comparison of CA 19-9 and carcinoembryonic antigen (CEA) levels in the serum of patients with colorectal diseases". In: *British journal of cancer* 49.2 (1984), p. 135.

[14] Feng Li, Teruyuki Kishida, and Masafumi Kobayashi. "Serum iron and ferritin levels in patients with colorectal cancer in relation to the size, site, and disease stage of cancer". In: *Journal of gastroenterology* 34.2 (1999), pp. 195–199.

[15] G Lindmark et al. "Limited clinical significance of the serum tumour marker Ca 72-4 in colorectal cancer." In: *Anticancer research* 16.2 (1996), pp. 895–898.

[16] OC Lunde and O Havig. "Clinical significance of carcinoembryonic antigen (CEA) in patients with adenocarcinoma in colon and rectum." In: *Acta Chirur gica Scandinavica* 148.2 (1982), pp. 189–193.

[17] A Mangano et al. "Complelent and Its Fractions (C3-C4) Pattern in Subjects with Msoplasia". In: *Journal of immunopharmacology* 6.3 (1984), pp. 147-162.

[18] Gerard Milano et al. "Serum prealbumin, retinol-binding protein, transferrin, and albumin levels in patients with large bowel cancer". In: *Journal of the National Cancer Institute* 61.3 (1978), pp. 687–691.

[19] L Molnar et al. "Correlation between the results of carcinoembryonal antigen (CEA) test and the clinical stage of colorectal carcinoma." In: *Acta chirurgica Hungarica* 27.1 (1986), pp. 27–34.

[20] Seung-Jae Myung. "Colon tumor and inflammation: is C-reactive protein possible colon tumor marker?" In: *The Korean Journal of Gastroenterology* 51.4 (2008), pp. 265–268.

[21] Tohru Nakagoe et al. "Prognostic value of carcinoembryonic antigen (CEA) in tumor tissue of patients with colorectal cancer." In: *Anticancer research* 21.4B (2001), pp. 3031–3036.

[22] Shivananda B Nayak et al. "Copper and ceruloplasmin status in serum of prostate and colon cancer patients". In: *Indian journal of physiology and pharmacology* 47.1 (2003), pp. 108–110.

[23] Richard L Nelson. "Iron and colorectal cancer risk: human studies". In: *Nutrition reviews* 59.5 (2001), pp. 140-148.

[24] Tadahiro Nozoe et al. "Increase in both CEA and CA19-9 in sera is an inde pendent prognostic indicator in colorectal carcinoma". In: *Journal of surgical oncology* 94.2 (2006), pp. 132–137.

[25] Seung-Yeol Park et al. "N-glycosylation status of B-haptoglobin in sera of patients with colon cancer, chronic inflammatory diseases and normal subjects". In: *International journal of cancer* 126.1 (2010), pp. 142–155.

[26] Bo E Persson et al. "A clinical study of CA-50 as a tumour marker for monitoring of colorectal cancer". In: *Medical oncology and tumor pharmacotherapy* 5.3 (1988), p. 165.

[27] Maja Prutki et al. "Altered iron metabolism, transferrin receptor 1 and ferritin in patients with colon cancer". In: *Cancer letters* 238.2 (2006), pp. 188–196.

[28] Henning Putzki et al. “Comparison of the tumor markers CEA, TPA, and CA 19-9 in colorectal carcinoma”. In: *Cancer* 59.2 (1987), pp. 223–226.

[29] Armin Quentmeier et al. “Carcinoembryonic antigen, CA 19-9, and CA 125 in normal and carcinomatous human colorectal tissue”. In: *Cancer* 60.9 (1987), pp. 2261–2266.

[30] M Rosandić-Pilas et al. “Relationship between tissue and serum concentrations of carcinoembryonic antigen (CEA) in gastric and colonic carcinomas.” In: *Acta medica Austriaca* 17.5 (1990), pp. 89-93.

[31] YT Van der Schouw et al. “Comparison of four serum tumour markers in the diagnosis of colorectal carcinoma”. In: *British journal of cancer* 66.1 (1992), p. 148.

[32] Jian-qiu Sheng et al. “Transferrin dipstick as a potential novel test for colon cancer screening: a comparative study with immuno fecal occult blood test”. In: *Cancer Epidemiology and Prevention Biomarkers* 18.8 (2009), pp. 2182–2185.

[33] Colin Walker and Bruce N Gray. “Acute-phase reactant proteins and carcinoem bryonic antigen in cancer of the colon and rectum”. In: *Cancer* 52.1 (1983), pp. 150–154.

[34] V Yasasever et al. “Serum values of CA72. 4 in patients with gastrointestinal system tumors comparison with CEA and CA 19.9.” In: *European journal of gynaecological oncology* 13.5 (1992), pp. 403–408.

Received January, 2019