Mapping Health Trajectories on Self Organizing Maps using COVID-19 Patient's Blood Tests

1st Carlos Arias-Alcaide
Dept Signal Theory and Communications
Universidad Rey Juan Carlos
Madrid, Spain
carlos.arias.alcaide@urjc.es

2nd Cristina Soguero-Ruiz
Dept Signal Theory and Communications
Universidad Rey Juan Carlos
Madrid, Spain
cristina.soguero@urjc.es

3rd Paloma Santos-Alvarez
Hospital Assistance Director
Hospital Universitario del Sureste
Madrid, Spain
paloma.santos@salud.madrid.org

4th Adrián García-Romero
Dept Management Control
Hospital Universitario del Sureste
Madrid, Spain
adrian.garcia@salud.madrid.org

5th Inmaculada Mora-Jiménez
Dept Signal Theory and Communications
Universidad Rey Juan Carlos
Madrid, Spain
inmaculada.mora@urjc.es

2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 978-1-6654-0126-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/BIBM52615.2021.9669706

***Abstract*—Since COVID-19 appeared in December 2019, scientists are researching new ways to improve the management of the disease. Considering machine learning approaches have proven to be very useful tools to discover hidden patterns in data, we propose in this paper to apply a Self Organizing Map (SOM) to characterize the health-status evolution of COVID-19 patients. The SOM is a neural network whose neurons can be represented as cells in a bi-dimensional grid preserving the mapping from the original space to the map units. We consider real-world data of hospitalized COVID-19 patients in a Spanish hospital during the first wave of the pandemic. Patients are represented by six blood tests (leukocytes and D-dimer, among others) in a daily basis. Besides, each patient is associated with one of two different health-status: favorable evolution (discharged home) and unfavorable evolution (exitus or admission to the intensive care unit). We show the potential of our approach by detailing the mapping of the health trajectory associated with different particular cases and drawing their trajectory on the bi-dimensional map of the SOM.**

***Index Terms*—SOM, machine learning, COVID-19, disease progression, blood biomarker.**

I. Introduction

The first case affected by *Severe Acute Respiratory Syndrome Coronavirus 2* outside of China was registered in January 2020 [1]. Rapidly, numerous people around the world started communicating symptoms related to this coronavirus. A new coronavirus disease named COVID-19 had started to quickly spread, producing an epidemic with important health and economic consequences [2]. Since that moment, the scientific community is doing research to improve the management of the COVID-19 disease. In this scenario, the design of models through machine learning (ML) approaches can provide with advanced techniques to extract relevant knowledge from clinical data. However, because of the lack of clinical protocols in the pandemic outbreak, data were not usually collected in a controlled manner. Both pharmacological treatments and the results of blood tests stand out among those data collected more regularly. Owing to the fast changes in the drug treatments during the first wave of the pandemic [3], produced by the insufficient number of studies, we focus here in the analysis of blood tests. In fact, blood tests continue being for practitioners the main source to know the patient health status [4] and make a decision about their treatment.

To contribute to the COVID-19 research from a data analytic perspective, we propose in this paper to use ML approaches, specifically artificial neural networks driven by the results of blood tests. Our goal is to gain knowledge about the health-status evolution (related to the disease progression) of patients hospitalized with COVID-19. Specifically, we propose the use of blood tests to build a Self Organizing Map (SOM) [5] where each neuron is potentially associated with a health-status. On the one hand, patients with Favorable Evolution (FE) are those who are discharged home with no previous admission in the Intensive Care Unit (ICU). On the other hand, we consider patients with Unfavorable Evolution (UE) as those either admitted to the ICU or who died (exitus, non survival) with no ICU admission. Medical practitioners have considered appropriate to include the ICU patients in the group of UE patients (regardless of the outcome), since the ICU admission is caused by a poor health status. In this paper, we call as "event day" the medical discharge day for patients with FE, and the day of ICU admission or exitus for patients with UE.

It is important to remark here that the SOM is characterized because it preserves the mapping from the high dimensional space to the map units [6]. Therefore, by identifying neurons with health-statuses, it is possible to use the SOM topology to show the progression of the patient health-status as a tour on a bidimensional map, what we named health trajectory.

The rest of the paper is structured as follows. Section II presents a description of the dataset collected by the Hospital Universitario del Sureste (HUS, Madrid, Spain), and the pre-processing applied for using these data in the proposed approach. In Section III, we detail the procedure for the SOM training and its application on our data, presenting also the

model for mapping the health status evolution. In Section IV, we show the health trajectories of several patients when mapped on the SOM, together with their clinical interpretation. Conclusions and future work are presented in Section V.

II. Dataset and Pre-processing

We start this section by describing our dataset. Then, we explain the process to adequate the data to the SOM training and to characterize the disease progression.

A. Dataset Description

Data were collected during the first wave of the COVID-19 pandemic by the HUS, a Spanish public hospital encompassing 200,000 citizens. Patients considered in this work were hospitalized between March 3 and June 28. However, those transferred to another hospital were discarded because lack of access to data collected by other hospitals.

For the health trajectory characterization, our dataset is composed of patients who presented more than one laboratory result (each result in a different day) for each blood test. In particular, only the blood tests registered while the patient was hospitalized with a COVID-19 diagnosis have been considered. This time goes from the admission date to: i) the discharge date for patients with FE, and ii) either the ICU admission date or the exitus date for patients with UE.

Based on the clinical knowledge of practitioners and some works in the literature [4], the following collection of six blood biomarkers (features) have been considered:

- *Leukocytes* (F_1). They are blood cells whose function is to defend the body [7]. Their level increases in response to both viral and bacterial infections [8] [9].
- *Lymphocytes in* % (F_2). They are a type of leukocyte (F_1) which level decreases in cases of viral infections and malnutrition [7] [8] [9].
- *D-Dimer* (F_3). This is one of the protein fragments produced when blood clotting is activated [7]. In a normal health situation, values are almost undetectable [8] [9].
- *Lactate dehydrogenase* (F_4). It is an enzyme found in tissues of internal organs such as the heart, lungs or liver [9]. High levels are linked with organic damage [8].
- *Aspartate transaminase* (F_5). This is also an enzyme, normally presented in the liver an the heart cells, which a high level indicating damage in those two organs [8] [9].
- *C-Reactive Proteine* (F_6). It is a protein, made by the liver [9], which increases when there is an inflammatory process (in any part of the body).

With the above considerations, the total number of patients in the dataset was 367: 328 with FE and 39 with UE (12 patients in ICU and 27 non-survival patients). For features F_2, F_4 and F_6 and group of patients (FE and UE), a descriptive analysis of the first blood tests (label "First") and of those obtained in the nearest previous date to the "event day" (label "Last") is presented in Fig. 1. Note that the three features show differences in their distributions. It can be observed that F_2 has a very characteristic pattern, since the most unfavorable cases generally present a lower value, both in the first and in the last blood test. This result is in line with the clinical knowledge in case of viral infections. For F_4, values are usually lower for patients with FE. Regarding F_6, one of the inflammation and organic damage indicators, it also shows values consistent with the known effects of the virus. Indeed, values of F_6 are lower for patients with FE, showing remarkable differences between the first and the last blood tests also in FE patients.

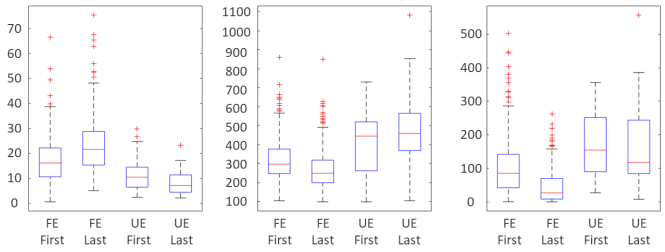

Fig. 1. Boxplot of the first and last blood test result for patients with FE (FE-First and FE-Last) and for those with UE (UE-First and UE-Last). Each panel shows a blood test: F_2 (left), F_4 (middle) and F_6 (right).

Apart from the high imbalance in the number of patients with FE and UE, our analysis revealed that there were also differences in their age distributions. To avoid that this skew is transformed in a bias in the model learned when using ML approaches [10], we apply here an undersampling strategy on the set of patients with FE. The undersampling was randomly performed so that the age distribution of the overrepresented group (those with FE) is similar to that of the underrepresented group (patients with UE). As a consequence, the final dataset resulted in 78 patients (39/39 patients with FE/UE).

B. Data Preparation

We explain here the two techniques applied on the final dataset: imputation and normalization. They are necessary for both training the neural network and for a subsequent mapping of the patient's health trajectory.

As previously mentioned, there was no well-defined medical protocol in the early months of the COVID-19 pandemic and some of the blood tests in Subsection II-A were not routinely collected. Although laboratory testing was recommended with a periodicity of 48-72 hours for stable patients, which might be reduced according to the patient's health progression [11], it is necessary to have one value per day and feature for mapping the health status on a daily basis.

In the final dataset, blood tests F_1 and F_2 (usually jointly provided) have the highest percentage of registered data: about 45.8% of days before the "event day", i.e. almost every 48 hours. Very similar percentages are obtained for F_5 and F_6, with 44.9% and 44.7% of the days, respectively. Finally, F_4 and F_3 covered the lowest rate of daily blood test values (40% and 37.9%), maybe because they were incorporated into the HUS protocol at the end of the considered period of time.

According to the previous analysis, we decided to impute missing values [12] to work with data on a daily basis. For the imputation, we applied the "last observation carried forward" [12] strategy, commonly used in the clinical literature [13] because of its simplicity. Similarly to a scenario

in which clinicians makes their decisions based on the last results, Fig. 2 illustrates the imputation process associated with six different blood test of a particular patient. Note that each blood test can be taken in a different day. For each day d with no new blood test result, the value registered for the same patient and blood test in the nearest previous day (circle) is copied (diamond) in the day d. For this particular case, our approach would allow us to analyze the health trajectory from 06/Apr (left vertical dotted line) to 12/Apr (right dotted line), since this time interval has values for the six blood tests.

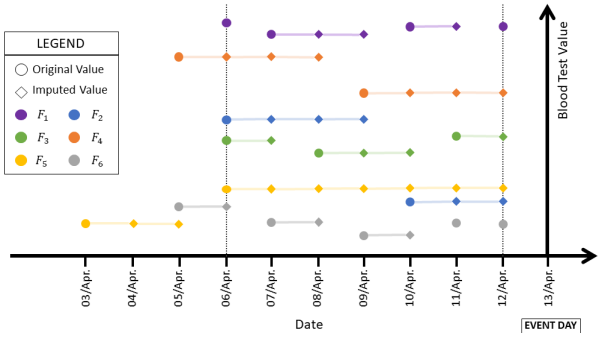

Fig. 2. Scheme for imputing on a daily basis the six blood tests (F_1 to F_6, in colour) for a particular patient. Original values are depicted as circles, while imputed values are diamonds. Horizontal segments represent the carry forward imputation process. Vertical dotted lines indicate the date of the first and last day when all tests have a value (original or imputed).

It is a common practice in ML to normalize numerical features [14] to avoid that those with higher dynamic range have more influence in the model. This is particularly important when considering distance-based algorithms such as SOM. The presence of outliers in the blood tests (see Fig. 1), suggests the use of a robust normalization procedure to scale the original features to a standard range. This is similar to the transformation using a sigmoid function for normalization [15], but with no strict saturation in the extremes, just to get the original values from the transformed ones. Though this recovery is not necessary in our approach, it may help the practitioner to analyse the patient's health-status evolution. fect the model used to map the trajectory.

Thus, for the i-th blood test F_i, we take the values of the closest day before the "event day" and compute both the median (Med_i) and the median absolute deviation about the median (MAD_i), according to [16]. Next, values of F_i in the interval $[\text{Med}_i - k \cdot \text{MAD}_i, \text{Med}_i + k \cdot \text{MAD}_i]$ are shifted around zero and linearly scaled using $F_i' = (F_i - \text{Med}_i)\,/\text{MAD}_i$. We used $k = 4$ as a reasonable value, empirically chosen after examining the feature distributions. To mitigate the influence of outliers, and trying to approximate a sigmoid function using a piecewise linear transformation, the values of F_i outside the previous interval are quasi-saturated using a linear transformation with a very small slope in the extremes.

III. Self Organizing Map

We start this section with a brief explanation about SOM and its training process with the data presented in Section II. Then, we interpret the results from a clinical viewpoint.

A. SOM Explanation and Model Training

The SOM is a neural network proposed by Kohonen [5]. Its design is based on statistical learning principles from a set $\mathcal{X}$ of N observations, i.e. $\mathcal{X} = \left\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, ..., \mathbf{x}^{(N)}\right\}$. The i-th observation is expressed by a D-dimensional vector such as $\mathbf{x}^{(i)} = \left[x_1^{(i)}, x_2^{(i)}, ..., x_D^{(i)}\right]$, with $D = 6$ (number of features) in this work. Regardless of the value of D, the SOM can be represented by a bidimensional grid of neurons (also named cells). A very important aspect of SOM refers to its topological properties, since similar observations in the original space are positioned in the same cell or in two very near cells of the map. This property is very important in our particular scenario, facilitating the interpretation of the patient's health trajectory. It is important to remark here that the SOM training is unsupervised, i.e., no information about the health-status of the patient is considered during training.

The size of the SOM (number of neurons G in the grid) must be established before training, since each observation in $\mathcal{X}$ is assigned to the closest neuron in the feature space. Though the most appropriate value depends on the specific scenario and dataset, authors in [17] propose $G \approx \sqrt{N/2}$ as a rule of thumb. As our dataset is composed of $N = 78$ patients, we considered $G = 6$ cells. Each cell is characterized by a representative vector also named codebook vector $\mathbf{c}^{(j)}$, with $j = 1, \ldots, G$. The codebook values are obtained by averaging features of the observations associated with the corresponding neuron. Thus, the design of the SOM is an iterative process where observations are assigned to one or other of the codebook vectors, which values are updated for T iterations. As proposed in [18], the number of iterations is determined as $T \approx 500G$, resulting in $3,000$ iterations.

For each iteration t, an observation $\mathbf{x}^{(i)}(t) \in \mathcal{X}$ is randomly selected, and the Euclidean distance to every codebook vector is computed. Then, the nearest codebook (e.g. $\mathbf{c}^{(b)}(t)$), called the Best Matching Unit (BMU), is identified and its features are updated by also considering features of the neighbouring codebooks. The nearest neurons to the BMU are more influential in the change than those which are further. This process is modeled by the neighbourhood function $h_{bj}(t) = \exp\left(-\frac{||\mathbf{r}_b - \mathbf{r}_j||^2}{2\sigma^2(t)}\right)$, where $\mathbf{r}_b$ and $\mathbf{r}_j$ are the positions of the nodes in $\mathbb{R}^2$, and $\sigma(t)$ is the width of the neighbourhood function, which decreases monotonically with the number of iterations. In [18], Kohonen also recommends to start with, at least, half the diameter of the network. This way, the codebooks update goes from a global environment (initial iterations) to a local one (final iterations, since σ is progressively reduced). Thus, the codebooks are updated as

$$\mathbf{c}^{(j)}(t+1) = \mathbf{c}^{(j)}(t) + \alpha(t) h_{bj}(t) \left[\mathbf{x}^{(i)}(t) - \mathbf{c}^{(j)}(t)\right] \; \forall j = 1, ...G$$

where $\alpha(t)$ is the learning rate controlling the difference between two consecutive updates of the same codebook. As indicated in [19], the learning rate should be also a decreasing function with the number of iterations.

To characterize the health progression, we considered for the SOM training the collection of $N = 78$ observations that,

potentially, more discriminate the patient's health condition. Thus, just the feature values associated with one day before the "event day" have been considered in training.

B. Interpretation of the Trained Model

The bidimensional grid of the SOM and the number and distribution of FE/UE patients per neuron after training is shown in Fig. 3. Though no information about the health status has been provided during training, note that patients assigned to neurons 1 and 6 (located far away in the grid) have the opposite health status (but homogeneous in the same neuron, "pure" neurons). The rest of neurons have assigned patients with both health statuses (non-pure neuron). Thus, neuron 2 is mostly represented by patients with FE (64% of patients), whilst neurons 3, 4 and 5 have a higher proportion of patients with UE ($61,5\%$, $81,8\%$ and $66,6\%$, respectively).

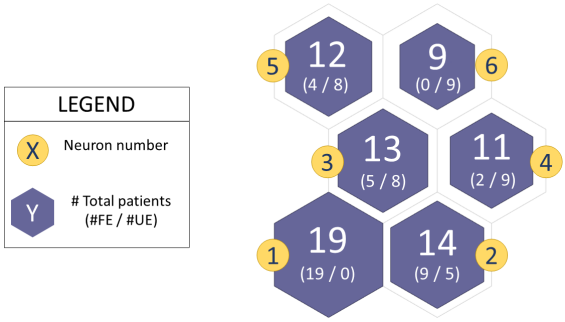

Fig. 3. SOM topology, with the neuron number embedded in a red circle. The total number of patients associated with each neuron and the distribution between FE and UE are indicated inside the neuron.

To interpret each neuron from a clinical perspective, we present in Fig. 4 the numerical value of each feature in the codebook vector (circles in black), with one panel per neuron. The 75% empirical confidence interval (CI) for each feature and health status, obtained from patients assigned to each neuron, is also represented: the CI depicted in blue/red is associated with FE/UE patients. For each neuron and feature, the median is represented as a dot inside the corresponding CI. Next, we provide a detailed analysis for each neuron.

As previously indicated, since neurons 1 and 6 just encompass patients with the same health status, values of the codebook features are inside the CI (see Fig. 4). Note, however, that mean (codebook) and median (marked inside the CI) might not coincide. Though just the normalized features are presented here, we have checked that de-normalized values are in accordance with what is expected in the literature [4]. This means that patients with FE in neuron 1 have moderate values of F_1' and high for F_2', while the opposite happens for patients with UE in neuron 6. With respect to F_3', F_4' and F_6', the values for UE patients (neuron 6) are higher than those for FE patients (neuron 1). There is not a significant difference for both groups of patients when considering F_5'.

With regard to neuron 5, the analysis of the CI reveals that both groups of patients show similar values in the blood test associated with F_1', F_3' and F_6'. However, the CI of both groups

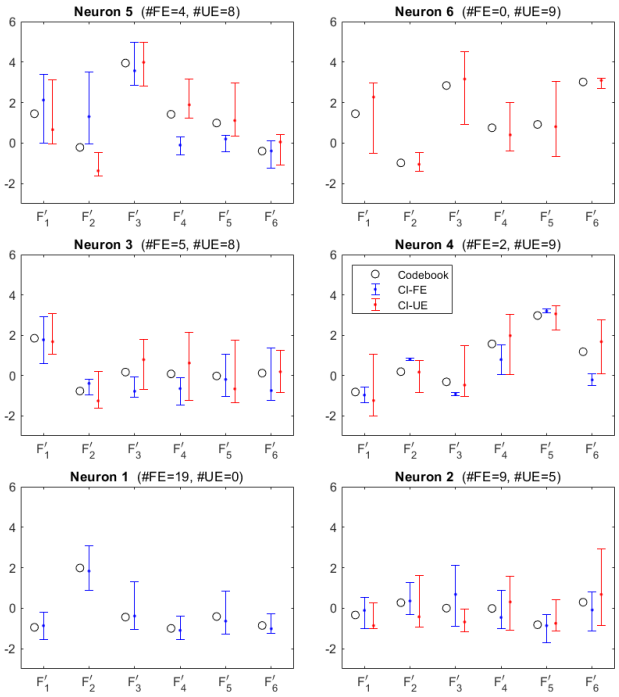

Fig. 4. Statistical description of the normalized features associated with each neuron after the SOM training. The arrangement of neurons (one per panel) is the same as in Fig. 3. The numerical value of each feature in the codebook vector is represented as a black circle. The 75% empirical CI associated with FE patients are in blue, while those of UE patients are in red.

is different for F_2', F_4' and F_5', suggesting that these features have been very decisive for prognosis.

As for neuron 4, note that the number of patients with FE is only two. Therefore, it does not seem adequate to draw conclusive conclusions about differences in the values of features in both groups of patients. Despite the above, note that the CI of both groups do not overlap when considering F_6'. Associated also with the reduced number of patients with FE, is the dispersion of values in features F_1', F_2', F_3' and F_5', which is higher for the UE patients than for the FE ones.

Regarding neuron 3, although features F_3' to F_6' seem to have similar values to those associated with a FE health status (as in pure neuron 1), most of the patients assigned to neuron 3 have an UE, likely due to the values of F_1' and F_2'.

Finally, note that the CI of features linked to patients encompassed by neuron 2 are similar for both groups. Note also that the CI associated with most of the features of patients with FE are similar to those of neuron 1. This might justify why both neurons are closely located in the topological grid.

IV. Mapping the Patient's Health Trajectory

We present in this section how the SOM can be used to map the patient's health progression as a trajectory in a two-dimensional grid. For this purpose, we take into account the blood tests on a daily basis, starting for each patient from the date with original or imputed values for all the considered blood test. Thus, for each day d and patient i, the observation

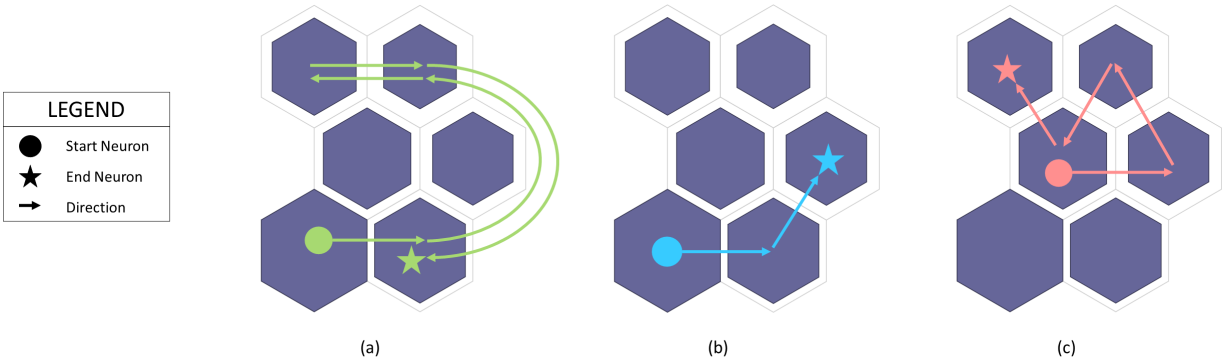

Fig. 5. Mapping of the health trajectory for each patient in Table I: (a) FE, (b) UE (ICU) and (c) UE (exitus before ICU admission). The beginning of the path is marked with the circle, while the star is placed on the neuron linked to the day before the event.

vector $\mathbf{x}^{(i,d)} = \left[x_1^{(i,d)}, x_2^{(i,d)}, ..., x_6^{(i,d)}\right]$ is constructed, and the normalization is performed. Then, the closest codebook vector to $\mathbf{x}^{(i,d)}$ is found, i.e. the BMU, and $\mathbf{x}^{(i,d)}$ is assigned to it. By considering consecutive days and creating the list of corresponding BMUs, the health trajectory can be drawn on the grid. Note that most of the values used to represent the trajectory have not been used in training.

For interpreting the health trajectory, we present in this section three representative cases. For each case, we show: (i) a list in Table I with the date, the values of the non-normalized features (F_i) and the neuron number (#Neur) the observation is assigned to; (ii) a sketch in of the grid (see Fig. 5) and the neurons where the observations are associated with as time evolves, going for each case from the first date in the list (circle in the graph) to the last one (star).

The first part of Table I and Fig. 5 (a) show the case of a patient with FE. From 21/Mar to 27/Mar, observations are in neuron 1: values of all blood tests, excepting F_3 and F_4 which are a bit high, are in the range associated with FE patients. From 28 to 31 March, the patient's health status changes to neuron 2: F_1 increases though it continues within values associated with FE patients, F_2 is halved and F_3 is almost doubled, moreover, there is a slight increment of F_4 and F_6. The next change in the trajectory, to neuron 6, is justified by an increase in F_1, a decrease in F_2 and, specially, a high change in F_3 (the value is tripled). The decrease of values in F_4 and F_6, could be the reason for a better health status on 04/Apr, changing from neuron 6 to neuron 5. The change from neuron 6 to neuron 2 in the last days of the trajectory is due to the progressive increase of F_2, while F_1, F_3 and F_6 decrease. It is interesting to remark that the last neuron in every trajectory is linked to the day before the "event day". It may therefore be possible that the favourable evolution when the patient is discharged home, could lead the observation to neuron 1. One might think that the final health status is worse than the initial one, since neuron 2 encompasses patients from both groups, while neuron 1 only have patients with FE. However, it must be taken into account that the first day presented in Table I and drew in the trajectory corresponds to the first day during hospitalization with values for the six blood tests. In other words, the patient may have been hospitalized several days before the first date indicated in Table I, since one or more of the six blood tests have not been taken. Thus, the first days of the hospitalization could have not been considered.

TABLE I
HEALTH TRAJECTORIES FOR THREE DIFFERENT PATIENTS: WITH FE, UE (ICU ADMISSION) AND UE (EXITUS BEFORE ICU ADMISSION)

Patient	Date	F_1	F_2	F_3	F_4	F_5	F_6	# Neur.
FE	21/Mar.	3.24	21.3	950	304	35	31.2	1
	22/Mar.	3.24	21.3	950	304	35	31.2	1
	23/Mar.	4.03	24.6	2760	467	58	67.4	1
	24/Mar.	4.93	24.6	2760	467	58	67.4	1
	25/Mar.	5.23	25	760	542	51	79.7	1
	26/Mar.	5.23	25	760	542	51	79.7	1
	27/Mar.	5.23	25	760	542	51	79.7	1
	28/Mar.	6.59	11.8	1370	555	38	196.1	2
	29/Mar.	6.59	11.8	1370	555	38	196.1	2
	30/Mar.	7.64	12.6	1370	502	28	249.6	2
	31/Mar.	7.64	12.6	1370	502	28	249.6	2
	01/Apr.	9.02	10.9	4550	410	27	259.9	6
	02/Apr.	9.02	10.9	4550	410	27	259.9	6
	03/Apr.	9.02	10.9	4550	410	27	259.9	6
	04/Apr.	9.41	10.8	4570	262	24	100	5
	05/Apr.	9.41	10.8	4570	262	24	100	5
	06/Apr.	9.41	10.8	4020	345	23	167.9	6
	07/Apr.	9.41	10.8	4020	345	23	167.9	6
	08/Apr.	8.73	14.7	4020	331	29	165.1	2
	09/Apr.	8.73	14.7	4020	331	29	165.1	2
	10/Apr.	7.95	18.4	2400	242	25	115.2	2
UE (ICU admis.)	23/Mar.	4.04	22.1	590	397	55	58.9	1
	24/Mar.	3.97	13.8	460	365	42	81.9	2
	25/Mar.	3.97	13.8	460	365	42	81.9	2
	26/Mar.	6.74	6	540	497	77	139.7	4
	27/Mar.	7.97	5.7	580	566	61	194.1	4
UE (Exitus before ICU admis.)	01/Apr.	15.55	8.3	890	526	63	108.6	3
	02/Apr.	15.55	8.3	890	526	63	108.6	3
	03/Apr.	12.06	9.6	1100	509	94	208.5	4
	04/Apr.	12.06	9.6	1100	509	94	208.5	4
	05/Apr.	11.1	6.9	2050	390	106	145.8	4
	06/Apr.	11.1	6.9	2050	390	106	145.8	4
	07/Apr.	26.46	2.1	2670	429	49	217.3	6
	08/Apr.	26.22	2	2180	382	61	76.5	3
	09/Apr.	26.22	2	2180	382	61	76.5	3
	10/Apr.	20.73	4.3	4510	506	139	18.9	5
	11/Apr.	20.73	4.3	4510	506	139	18.9	5
	12/Apr.	20.73	4.3	4510	506	139	18.9	5

The health evolution of a patient with UE (ICU admission) is presented in Table I and Fig. 5 (b). Note that on 23/Mar the observation vector associated with the patient's health status

is in neuron 1 (just with patients with FE). On 24/Mar, values of F_1 and F_2 decrease while that of F_6 increases, leading the change to neuron 2. On 26/Mar there is another considerable decrease in F_2 and a substantial increase in F_1 and F_6. It is also accompanied by an increase in F_3, F_4 and F_5, which produces the movement from neuron 1 to neuron 4.

The third case corresponds to an UE patient (exitus before ICU admission). As indicated in the final part of Table I, the first day in the trajectory (01/Apr) starts in neuron 3: F_1 is above 10, F_2 is low compared to standard blood test values, and the other four features have high values. On the 03/Apr, F_1 decreases and F_2 and F_3 increase while F_6 doubles. This moves the observation of the patient's health status from neuron 3 to neuron 4, which is a worse health status. Though F_2 decreases and F_3 doubles on 05/Apr, observations remains in neuron 4, most likely because the decrease on values of F_4 and F_6. On 07/Apr, the patient's health worsens even more: F_1 doubles again, F_2 is 2, F_3 and F_6 also increase. This blood test results make the observation vector being assigned to neuron 6. On the day 08/Apr, the patient's health status seems to worsen because, even when the value of F_6 decrease considerably, making the observation vector changes to neuron 3, the rest are very similar to the previous day. The last blood test of the patient before the "event day", registered on 10/Apr, indicates a health status in accordance with neuron 6: very high levels at F_1, blood tests F_2, F_3, F_4 and F_6 very low and F_5 very high too. The trajectory is depicted in Fig. 5 (c).

V. Conclusions and Future Work

The use of the SOM as a tool for characterizing the patient's health progression has been validated. For this purpose, we have considered a real-world dataset composed of blood tests associated with hospitalized COVID-19 patients. Special attention has been paid to consider patients with similar demographic characteristics. In spite of the reduction performed in the number of patients to deal with the imbalance problem, what limits the grid size, this paper shows promising results to use the SOM in the clinical setting. We consider that the description of the patient's health trajectory, both on the map and in relation with the values of the patient's features, is a valuable tool to gain intuition about the patient's health progression. Our work is a first approach which could be considered as a proof of concept, with many potential extensions even to other scenarios.

Several lines are opened as future work. Firstly, we propose to apply oversampling of the minority group instead of undersampling of the majority one. By increasing the size of the training observations, e.g. by applying the SMOTE approach, it is expected that the grid size (number of neurons) increases, leading to more specific neurons from a clinical viewpoint. As a consequence, it is likely that observations can change the associated neuron even in consecutive days, potentially leading to highly diverse health trajectories.

In this work we have considered the majority group as that identifying the "representative" health status in a neuron. Our second line of work is to perform a probabilistic characterization of the health status linked to each neuron. In particular, we propose to design a *Maximum A Posteriori* classifier [15] per neuron by using the Naïve-Bayes technique to capture the diversity in non-pure neurons.

The successful implementation of the previous lines of work would lead to a design discriminating several groups of health status: home-discharged patients and non-survival patients, in both cases with and without ICU admission, separately. The consideration of additional features (e.g. age, sex, drug treatment, vital signs or even vaccination events) could also provide very useful insights to support clinical decisions.

Acknowledgment

This work has been partly supported by the Spanish Research Projects PID2019-106623RB-C41 and PID2019-107768RA-I00, by the Project ReCOVID (M2408, COVID-19 call funded by Universidad Rey Juan Carlos), as well as by the Community of Madrid in the framework "Encouragement of Young Phd students investigation" (Mapping-UCI).

References

[1] F. Pinotti and et al., "Tracing and analysis of 288 early SARS-CoV-2 infections outside China: A modeling study," *PLOS Medicine*, vol. 17, no. 7, 2020.

[2] Y. Li and J. E. Mutchler, "Older adults and the economic impact of the covid-19 pandemic," *Journal of Aging & Social Policy*, vol. 32, no. 4-5, pp. 477–487, 2020.

[3] W. Moumouh-Ait, M. Sevilla-García, A. García-Romero, C. Soguero-Ruiz, and I. Mora-Jiménez, "On the statistical differences in the pharmacological treatment of covid-19 patients," in *Proc of the IEEE EMBS Intl Conf on Biomedical and Health Informatics*, 2021.

[4] I. T. Parsons and et al., "The use of routine blood tests to assist the diagnosis of covid-19 in symptomatic hospitalized patients," *Annals of Clinical Biochemistry*, vol. 58, no. 4, pp. 318–326, 2021.

[5] T. Kohonen, "Self-organized formation of topologically correct feature maps," *Biological Cybernetics*, vol. 43, pp. 59–69, 1982.

[6] ——, "The self-organizing map," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464–1480, 1990.

[7] J. E. Hall and A. C. Guyton, *Textbook of medical physiology*. Saunders Elsevier, 2015.

[8] R. Hoffman, E. Benz, L. Silberstein, H. Heslop, J. Anastasi, and J. Weitz, *Hematology: Basic Principles and Practice*, ser. Churchill Livingstone. Saunders/Elsevier, 2013.

[9] D. Kasper, A. Fauci, S. Hauser, D. Longo, and J. Jameson, *Harrison's Principles of Internal Medicine*. McGraw-Hill Education, 2015, vol. 2.

[10] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, *Learning from Imbalanced Data Sets*. Springer, 2018.

[11] K. P. Eaton and et al., "Evidence-Based Guidelines to Eliminate Repetitive Laboratory Testing," *JAMA Internal Medicine*, vol. 177, no. 12, pp. 1833–1839, 12 2017.

[12] R. J. Little and D. B. Rubin, *Statistical Analysis with Missing Data*. John Wiley & Sons, 2019, vol. 793.

[13] C. Soguero-Ruiz and et al., "Data-driven temporal prediction of surgical site infection," in *AMIA Annual Symposium Proceedings*, vol. 2015, 2015, p. 1164.

[14] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, ser. Springer Series in Statistics. NY, USA: Springer, 2001.

[15] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer Science+Business Media LLC, 2006.

[16] R. Maronna, R. Martin, V. Yohai, and M. Salibián-Barrera, *Robust Statistics:theory and methods (with R)*. John Wiley & Sons, 2019.

[17] K. V. Mardia, J. T. Kent, and J. M. Bibby, *Multivariate Analysis*. Academic Press, 1995.

[18] T. Kohonen, *Self-Organizing Maps*. Springer, 2001.

[19] L. V. Fausett, *Fundamentals of neural networks: Architectures, algorithms, and applications*. Prentice-Hall, 1994.