Clustering Corticosteroids Responsiveness in Sepsis Patients using Game-Theoretic Rough Sets

Rahma Hellali, Zaineb Chelly Dagdia, Karine Zeitouni

Clustering Corticosteroids Responsiveness in Sepsis Patients using Game-Theoretic Rough Sets

Rahma Hellali
Université Paris-Saclay,
UVSQ, DAVID, France
Email: rahma.hellali@uvsq.fr

Zaineb Chelly Dagdia
Université Paris-Saclay,
UVSQ, DAVID, France
Université de Tunis,
Institut Supérieur de Gestion
de Tunis, LARODEC, Tunisia
Email: zaineb.chelly-dagdia@uvsq.fr

Karine Zeitouni
Université Paris-Saclay,
UVSQ, DAVID, France
Email: karine.zeitouni@uvsq.fr

***Abstract*—Performing data mining tasks in the medical domain poses a significant challenge, mainly due to the uncertainty present in patients' data, such as incompleteness or missingness. In this paper, we focus on the data mining task of clustering corticosteroid (CS) responsiveness in sepsis patients. We address the issue and challenge of missing data by applying Game-Theoretic Rough Sets (GTRS) as a three-way decision approach. Our study considers the APROCCHS cohort, comprising 1240 sepsis patients, provided by the Assistance Publique–Hôpitaux de Paris (AP-HP), France. Our experimental results on the APROCCHS cohort indicate that GTRS maintains the trade-off between accuracy and generality, demonstrating its effectiveness even when increasing the number of missing values.**

I. Introduction

Due to its high mortality, incidence, and morbidity, sepsis is regarded as one of the most serious diseases that impact people's lives. The Third International Consensus Definition for Sepsis and Septic Shock (Sepsis-3), in 2016, defined sepsis as a "life-threatening organ dysfunction resulting from dysregulated host responses to infection" [1]. Immunologically, the human body releases some immune chemicals into the blood to fight the encountered infection. These released substances cause extensive inflammation, resulting in blood clots and leaking blood vessels. As a consequence, blood flow is disrupted, depriving organs of nutrition and oxygen, and hence, resulting in organ damage. The Sequential Organ Failure Assessment (SOFA) score [2] is used to codify the degree of organ dysfunction. It is difficult to estimate the global burden of sepsis. The study conducted in [3], estimated that in 2017 there were 48.9 million cases and 11 million sepsis-related deaths all over the globe, which accounted for almost 20% of all global deaths. There is no current diagnostic test for sepsis.

Knowing that there are still no specific interventions to control immune responses to invading pathogens [4], for sepsis, researchers have looked at the biological underpinnings of sepsis to see if there are any treatments that could help. Because of their impact on the immune system, corticosteroids have received a lot of attention [5]. The hormonal route from the hypothalamic-pituitary gland to the adrenal glands promotes corticosteroid synthesis in sepsis [6], [7]. These hormones affect inflammation through the formation of white blood cells, cytokines, and nitric oxide. The timing of corticosteroid administration may be a key component in therapy response. Short-term mortality was found to be higher in observational studies when hydrocortisone was started later. It is expected that corticosteroid treatment is advantageous for sepsis patients for these reasons and that variations in dose, timing, or duration of corticosteroid treatment may alter the patient response to treatment differently [8].

This paper delves into the data mining task [9] of clustering corticosteroid (CS) responsiveness in sepsis patients using the APROCCHS cohort provided by Assistance Publique–Hôpitaux de Paris (AP-HP), France. The cohort includes 1240 sepsis patients. A key challenge in this task is the presence of missing data.

Grouping data with missing values is one of the primary difficulties in clustering. There are commonly two strategies to deal with missing values [10]. The first strategy is based on preprocessing techniques [11]. Generally, it adopts deleting the whole row containing missing values or replacing the missing values based on experts' rules [12]. Some common missing values imputation techniques include replacing missing values with the mean, median, or mode of the available data for that feature [13]. The hot Deck Imputation method is replacing missing values by randomly selecting a value from another similar data point in the same dataset [14]. Using the values of K-nearest neighbors in the feature space to estimate the missing value [15]. Linear Regression Imputation aims to predict the missing values using linear regression based on other variables in the dataset [16].

The process of filling in missing values can potentially introduce a significant amount of imputation bias and uncertainty. It is important to recognize that missing values can be informative and carry meaningful implications. In certain instances, the absence of data itself can convey valuable information or signify a particular category or state. Imputing these missing values may result in distorting the original meaning or introducing artificial patterns into the dataset. In such situations, it is advisable to treat the missing values as a distinct category or conduct a separate analysis specifically

on the subset of data that contains missing values. Also, imputation during preprocessing has reportedly been found to compromise the accuracy and consistency of classification outcomes. Thus, these methods are not recommended specifically when we deal with medical data because they can bias the medical results.

The second strategy relies on incorporating mechanisms in the clustering model [17] which means that we will not impute or preprocess the missing values, but the internal functioning of the algorithm will handle automatically the missing values. Examples of works belonging to this strategy are [18] which allocated missing value objects to a cluster with a large number of missing values and [19] which assigned objects containing missing data to clusters based on their neighbors.

Moreover, the theory of rough sets provides a valuable framework for analyzing incomplete information through the use of approximations [20]. This approach allows us to delve into the realms of uncertain and imprecise data, aiding in our understanding of complex systems. According to the research conducted by [21], the application of rough set theory has been observed across various fields and domains. In [22], authors integrated the Variable Precision Rough Set (VPRS) approach with Bayesian principles. In [23], the idea is to combine VPRS with fuzzy rough set methods to create flexible decision rules. In essence, both papers share a common objective of tackling information imprecision by employing probabilities (within the framework of VPRS) and fuzziness (which allows for handling partial matching of rules' antecedents). Their ultimate aim is to derive interpretable decision models from the available data. Authors in [24] introduced the Learn++.MF, an innovative ensemble-of-classifiers algorithm designed to address the challenge of missing features in supervised classification. It creates an ensemble of classifiers, each trained on a random subset of available features. When classifying instances with missing values, the algorithm employs majority voting from classifiers that were trained without the missing features. The study demonstrates that Learn++.MF effectively handles significant amounts of missing data, with only a gradual decline in performance as the missing data increases. In biomedicine and healthcare, rough set theory has been applied for disease diagnosis [25], medical image analysis [26], and patient profiling [27].

Focusing on the second strategy, mentioned above, and tackling the challenge of missing data in sepsis patients' records, we apply Game-Theoretic Rough Sets (GTRS) as a three-way decision approach. The aim is to assign patients with incomplete records to the appropriate clusters automatically. In order to study the efficiency of the algorithm application on our clinical data, we aim to answer the following research questions:

- **RQ1:** How can the percentage of missing values affect the performance of the algorithm?
- **RQ2:** Does the k nearest neighbors has an impact on the results?
- **RQ3:** Can the percentage of increasing and decreasing initial values of α and β influence the results?

The rest of this paper is structured as follows. Section II presents the fundamentals of three-way decisions using GTRS. Section III details the application of GTRS, as a three-way decision approach for handling missing data, for clustering CS-responsiveness in sepsis patients. The experimental setup is introduced in Section IV. The results of the performance analysis are discussed in Section V, and conclusions are presented in Section VI.

II. Three-way decisions using rough sets

A. Three-way clustering

The theoretical foundation of three-way clustering is based on the theory of three-way decisions introduced by Yao [28]. Assuming the existence of a set $U = \{o_1, o_2, o_3, ...\}$ which is referred to as the universe of objects, a clustering method will produce a collection of sets $\{c_1, c_2, c_3, ...\}$, where each set c_k contains a group of objects belonging to that specific cluster. Every object o_i in the set has A attributes, represented as $o_i = (o_i^1, ..., o_i^A)$, with o_i^a indicating the value of the a^{th} attribute associated with the i^{th} object.

In traditional clustering, a cluster is usually represented by a single set, indicating that objects within the set definitely belong to a cluster and those outside the set definitely do not belong to it. In situations characterized by uncertainty and a lack of information, two-way decisions are not always feasible from a decision-making perspective, such as in the case of clustering.

A practical and reasonable alternative is to adopt a three-way decision approach, which introduces three options for decision-making, rather than the traditional binary choice. Specifically, we can decide whether an object belongs to a cluster, whether it does not belong to a cluster, or whether it is uncertain whether the object belongs to a cluster or not. This concept of three-way decision-making leads to what is known as three-way clustering.

To define three distinct regions - inside, partial, and outside - an approach involving an evaluation function and a set of thresholds can be employed. The evaluation function quantifies the association or correlation between an object and a cluster, while the thresholds set limits on this relationship for inclusion in each of the regions. Let $e(c_k, o_i)$ be an evaluation function that represents the association between a specific cluster c_k and an object o_i, and let (α, β) be a pair of thresholds. The three regions are defined as follows.

$$Inside(c_k) = \{o_i \in U | e(c_k, o_i) \geq \alpha\}, \tag{1}$$

$$Outside(c_k) = \{o_i \in U | e(c_k, o_i) \leq \beta\}, \tag{2}$$

$$Partial(c_k) = \{o_i \in U | \beta < e(c_k, o_i) < \alpha\}, \tag{3}$$

This means that when the evaluation of an object is equal or above the threshold α, it is considered to be part of the Inside(c_k) group. Conversely, if the evaluation is at or below the threshold β, the object is regarded as being in the Outside(c_k) group. If the object's evaluation falls between the two thresholds, it is included in the Partial(c_k) group. Thus,

inclusion in distinct regions is governed by the thresholds (α, β), and varying their settings results in different regions. The automatic determination of these thresholds is a crucial research topic in this context.

In this regard, and based on the work proposed in [29], we utilize the three-way framework to handle data with missing values which involves three steps. The overall functioning is presented in Figure 1. Initially, the set of objects U is partitioned into two sets: C and M. Set C comprises objects that have no missing data, while set M contains those that have missing values. Objects in set C are clustered using conventional algorithms, such as K-means [30], under the assumption that since these objects have no missing values, the level of uncertainty is low, and conventional approaches are more suitable for clustering such objects (Figure 1 (1)).

The second step (Figure 1 (2)) involves creating an incomplete data set from C while maintaining a similar rate of missing values to that of dataset U. For instance, if 30% of objects in the original dataset has missing values, approximately 30% of objects will be randomly chosen from C to induce missing values. This results in partitioning C into two additional sets: the constructed dataset comprising objects with missing values denoted as U_m, and the remaining objects in C with no missing values designated as U_c. This step assists in selecting appropriate values for (α, β) thresholds that will enable the clustering of objects with missing values.

The third step (Figure 1 (3)) involves determining the inclusion of objects with missing values, denoted as M, in the three-way framework. To employ three-way clustering on data with missing values, it is necessary to calculate the evaluation function $e(c_k, o_i)$, as specified in Equations 1, 2, and 3. This function measures the association between an object o_i and cluster c_k and can be defined in various ways. In our case, and as proposed in [29], we utilize an evaluation function that is based on the proportion of nearest neighbors for object o_i that belongs to cluster c_k:

$$e(c_k, o_i) = \frac{\text{Number of } o_i \text{ neighbors belonging to } c_k}{\text{Total neighbors of } o_i} \quad (4)$$

In order to determine the neighbors, a specific distance metric is required. For this example, we utilize the euclidean distance as follows:

$$d(i,j) = \sqrt{\sum_{a=1}^{A} \left(O_i^a - O_j^a\right)^2} \quad (5)$$

Here, o_i^a represents the value of the a^{th} attribute of the i^{th} object and any attributes with missing values are disregarded during distance computation. By utilizing the aforementioned distance metric, it is possible to calculate the distances of each o_i with missing values from all objects in U_c. After sorting these distances, the nearest neighbors for each o_i can be determined. Upon sorting these distances, the nearest neighbors can be identified. After determining the evaluation functions, Equations 1, 2, and 3 can be employed to determine the inclusion of objects into one of the three regions.

The goal of this approach is to enhance the clustering quality of data containing missing values. In this regard, two metrics need to be calculated based on the thresholds (α, β) as follows:

$$Accuracy(\alpha, \beta) = \frac{\text{Correctly clustered objects}}{\text{Total clustered objects}}, \quad (6)$$

$$Generality(\alpha, \beta) = \frac{\text{Total clustered objects}}{\text{Total objects in } U} \quad (7)$$

where *Accuracy* refers to how well we cluster objects with missing values, whereas *generality* refers to the fraction of objects that were clustered in the first place. Thus, as defined in [29], this goal can be approached from the perspective of a trade-off between accuracy and generality of the clustering.

B. Game theoretic rough sets

GTRS is based on a game-theoretic concept and formulation to estimate thresholds of the three-way decisions [31], [32]. The thresholds are interpreted based on a trade-off solution between numerous criteria used to analyze rough sets in a game scenario [33], [32]. Specifically, to increase the overall quality of three-way decisions, GTRS formulates strategies for players in the form of adjustments in thresholds. Each player contributes to the game by configuring the thresholds in order to optimize the game's benefits/rewards and utilities. The overall goal of a game in GTRS is to choose appropriate thresholds for three-way decisions with respect to the available criteria and presented information.

In GTRS (Figure 1 (4)), a typical game consists of three main elements: (i) game players, (ii) strategies, and (iii) payoff or utility functions. These components are usually defined as a tuple $\{P, S, u\}$, where [34]:

- **Game players:** The game players are denoted by a set P. The players in the game are selected to reflect the overall purpose of the game.
- **Strategies:** In the game, each player contributes by playing different strategies. The set of strategies available to player i is denoted by S_i. All possible strategy sets are denoted by the following Cartesian product: $S = S_1 \times S_2 \times \ldots \times S_n$, where S contains ordered pairs of the form $(s_1, s_2, \ldots, s_n)$ such that $s_1 \in S_1, s_2 \in S_2$ and $s_n \in S_n$. Each ordered pair in S is called a *strategy profile* and represents a certain situation encountered in a game.
- **Payoff functions or utility:** The payoff function, also called utility, for the players are defined via a set $u = (u_1, \ldots, u_n)$; where each u_i represents a real-valued utility function for player i and it maps the strategy profiles to real values ($u_i : S \mapsto \Re$). The payoffs reflect the utilities of performing or selecting a specific strategy.

Every player in a game seeks to execute a strategy that maximizes its payoff. The players' strategies, on the other hand, have an impact on their opponents' payoffs. The game solution is used to select a balanced and trade-off point based on all players' utilities. The *Nash equilibrium* is generally used to determine game solution or game outcome in GTRS.

Let us consider a strategy profile $s_{-i} = (s_1, s_2, \ldots, s_{i-1}, s_{i+1}, \ldots, s_n)$. s_{-i} is a strategy profile of all the players in the game except player i, and which can be further denoted as $s = (s_i, s_{-i})$. This means that all the players except i are committed to play s_{-i} and player i choosing s_i. The strategy profile $(s_1, s_2, \ldots, s_n) = (s_i, s_{-i})$ is a Nash equilibrium, when [35],

$$\forall i, \forall s_i^{'} \in S_i, u_i(s_i, s_{-i}) \geq u_i(s_i^{'}, s_{-i}), where(s_i^{'} \neq s_i) \quad (8)$$

This means that for all players i, their respective strategies, i.e., s_i is the best response to s_{-i}. In other words, a strategy profile constitutes a *Nash equilibrium* when no player is benefited from changing his/her strategy alone. The description presented above formulates a game in GTRS. It is to be noted that we may not be able to reach effective thresholds that meet the demands of the underlying applications with a single and non-repeated game. We, therefore, need to play the game several times; where in each play the goal is to keep modifying and refining the thresholds until we attain certain performance goals; e.g., a balance between accuracy and generality. The GTRS seeks an appropriate design of the threshold levels that are used in the three-way decisions framework, presented in Section II-A, by forming a game and applying concepts such as game solution and repeating games.

III. Application

A. Data Source

RECORDS[1] is a European research project that aims to quickly detect whether a patient is sensitive or resistant to the treatment of sepsis with corticosteroids. The project's clinical trial is an adaptive clinical trial that evaluates the efficacy of biomarkers and machine learning algorithms in defining patients' corticosteroid resistance, with the goal of optimizing their management. The project has adopted a distinctive approach to effectively analyze the severity of sepsis cases by collecting data on patients' demographics, health outcomes, and samples. This data collection has resulted in the creation of a first sepsis cohort, known as APPROCHS, which serves as an exceptional resource for medical research.

The paper considers the APROCCHS cohort that has been provided by the Assistance Publique–Hôpitaux de Paris (APHP) which is the university hospital trust operating in Paris, France, and its surroundings. It is the largest hospital system in Europe and one of the largest in the world. The goal of the cohort is to allow the investigation of qualitative interactions between clinical phenotypes and survival benefits or harms from corticosteroids (CS), i.e., to permit defining sensitivity and resistance to CS.

B. Data Description

The APROCCHS cohort gathers 1240 adult septic shock patients who are treated with or without CS. Each patient is characterized by 5645 features, also, called *risk factors* reflecting characteristics until Day 90.

Data were collected with a specification indicating whether the patients were treated with corticosteroid or with a placebo. A placebo is a substance or treatment that is given in the same manner as an active drug or treatment being tested but does not have any active ingredients or therapeutic effects [36].

C. Data Pre-processing

In this section, we explain the different data pre-processing tasks that we have performed on the APROCCHS cohort, namely: feature selection, data enrichment, data labeling and data cleaning.

1) Feature selection: Because sepsis is a time-sensitive disease, the likelihood of survival is significantly increased by early detection and treatment. This study focuses on using variables accessible at the earliest stage, especially at Day 0 of hospitalization, for predicting patients' responsiveness to corticotherapy in order to optimize accurate intervention. Specifically, from the initial pool of 5645 features –which reflects features from Day 0 until Day 90–, and by focusing only on features at Day 0, we were able to carefully choose a selection of 24 critical attributes following significant consultation with the respected medical specialists at APHP. This selection procedure entailed careful study and examination of each feature's relevance and significance in regard to our research with respect to the guidance and experience of the APHP healthcare experts.

The collected data is divided into two categories: static and dynamic. The first category includes information on the patient's current condition as well as personal information such as identification number, sex, weight, age, origin, date of hospitalization, and whether or not an antibiotic was administered before Day 0. These traits are noted at the time of admission and remain constant during hospitalization. The second category consists of dynamic elements that can be captured once or more times daily during hospitalization and are related to patient vital signs and laboratory testing. Admission type, infection date, infection place, and examination type are a few examples of dynamic characteristics that have only been once recorded. These data are often gathered before administering treatment. The sequential organ failure assessment (SOFA) score [37], ventilation, vasopressor use, and prescribed therapy dose are a few examples of characteristics that were recorded during the whole hospital stay and are associated with patients' responsiveness to treatment.

2) Data enrichment: Data enrichment relies on the process of adding new variables based on pre-existing ones in order to further explain the data and increase the precision of prediction algorithms. It improves detecting previously hidden relationships and patterns in the data. Following the guidelines of the APHP medical specialists, we generated the variable AR_INF_Type, which represents the source of infection and which was obtained from the diagnosis date and the hospital admission date variables. Furthermore, the values of the cortisol variable have been adjusted using a dataset that was given

[1]https://www.fhu-sepsis.uvsq.fr/rhu-records-4

by medical professionals and provides appropriate values for this characteristic. For a proper diagnosis and treatment plan, it is essential to know whether a patient received an antibiotic before being brought to the hospital. The medical staff's choice of the proper doses for the patient during their hospital stay will depend on this information, in addition to the machine learning model. A new feature entitled "ANTIBIOTIC" was created in order to acquire this information. A value of 1 of this characteristic implies that the patient took an antibiotic while a value of 0 means that he/she did not. The new variable "ANTIBIOTIC" comprises information on 690 patients who did not receive antibiotics prior to admission to the hospital and 550 patients who did.

3) Data labelling: For a patient who is enrolled in the study on Day 0, either corticosteroid medication or a placebo is administered every 4 to 6 hours while a number of features that indicate the patient's improvement are tracked. Daily feature values are recorded while each patient is observed for 90 days. The APHP healthcare experts have created precise standards for figuring out whether a patient would benefit from corticotherapy or not. Patients are specifically categorized as cortico-sensitive (i.e., responders) if all four of the following conditions are satisfied after 14 days of therapy:

- The patient survived.
- For at least 24 hours, there has been no vasopressor treatment.
- For at least 24 hours, the patient has been off of mechanical ventilation.
- The SOFA score is under 6.

The patient is generally considered cortico-resistant or a non-responder if the conditions are not satisfied, which is regarded a negative therapy response. As a result, the label is set to 1 or 0, indicating whether or not the patient reacted to the therapy on Day 14. Finally, patients who did not adhere to the aforementioned rule were eliminated from the cohort, leaving 1234 patients. This was done to preserve the integrity of our data and in accordance with the guidelines provided by medical specialists. The distribution of patients in the APPROCCHS cohort is shown in Table I.

4) Data cleaning: Particularly in important domains such as health, data cleaning and feature engineering are crucial steps in the data analysis process. These aspects have a significant impact on the decision-making process and the performance and accuracy of machine learning models. Dealing with the raw sepsis data that was gathered presented multiple challenges for this investigation. The APROCCHS cohort has a low rate of duplicate data, but in order to have accurate results with the three-way approach clustering, we have dropped duplicated patients. As a result, 1233 sepsis cases were still included in the cohort.

D. Three-way clustering with Game-theoretic rough sets

In this section, the application of the three-way clustering with GTRS, recently proposed in [29], is demonstrated using the pre-processed APROCCHS cohort. The objective is to cluster sepsis patients into two groups to reflect their responsiveness or not to CS while the model internally handles missing values.

1) Missing data description and handling: The only data pre-processing step that was not applied so far to the APROCCHS cohort is the task of handling missing values. As previously mentioned in Section I, the fact of imputing (replacing) or deleting the tuples containing missing values may significantly influence the conclusions drawn from the applied data mining task; specifically when it comes to a sensitive and critical domain as such is the medical domain. Pre-processing missing values may jeopardize the quality and reliability of the machine learning results; which is in our case the clustering task. As mentioned in Section I, a more appropriate and suitable strategy, to handle missing values, is to equip the clustering model with a mechanism able to handle data with missing values. In our study, this will be achieved by applying GTRS for three-way clustering.

However, it is still important to mention that, with respect to the medical experts' guidelines, some missing values had to be filled based on the following received recommendations:

- *Risk factors which are tied to the vasopressor treatment, life status, mechanical ventilation, and SOFA score:* Replace the missing value found at Day_i using the same non-missing value which is registered at Day_{i-1}. This is explained by the fact that if the value has not been registered at Day_i then this means that there has been no change in the patient's risk factor at Day_{i-1}.
- *The label:* To ensure the data's integrity and in accordance with the guidelines of medical experts, some patients have been updated from cortico-sensitive to cortico-resistant.

By applying these guidelines, the APROCCHS cohort still witnesses some missing values. These are distributed over 7 risk factors which are tied to the *KNAUS* score indicating the impact of a disease (i.e., sepsis) on the patient's activities, the *MACCABE* score indicating the presence of an additional fatal disease and its severity, the *SOFA* score in the last 3 hours after admission to intensive care, the *body temperature* at the entrance to the unit of intensive care, the *severity index*, the *glycemic index*, and the *blood lactate level*.

These will be taken care of at the GTRS clustering model level instead of modifying the data itself, i.e., will neither be imputed nor deleted; as will be explained in the next sections.

2) Game formalization: As described in Section II-B, the players, the strategies, and the payoff or utility functions, are the three components which are needed to be defined to analyze problems with GTRS. The game formalization is as follows:

a) The objective of the game: The aim of this game is to improve the clustering performance of datasets with missing values. As stated in [29], this objective can be achieved by balancing the accuracy and generality of the clustering, as described by Equations 6 and 7.

b) The players: The game's ultimate objective and goal should be reflected in the players. In this regard, the players in this game present the clustering's accuracy and generality

TABLE I
DISTRIBUTION OF PATIENTS IN APPROCCHS

Cohort	APPROCCHS			
Group	Features	Sensitive/Improved	Resistant/Not improved	Total
Corticosteroid	5645	233	379	612
Placebo		213	409	622
Total	5645	446	788	1234
Characteristic	APPROCHS randomized controlled trial			

features. Let A denote player *Accuracy* and let G denote the player *Generality*. P = {A, G} represents the player's set.

c) The strategies: The strategies denote the different actions that a player can take in a game. To maximize her/his rewards/benefits, each player adopts a strategy. As demonstrated in [29], when different thresholds are used in the game, the properties of accuracy and generality are influenced differently. Consequently, changes and variations in thresholds can be considered as feasible strategies. Three strategies are considered in our context:

- Decreasing the threshold α — defined as $(\alpha \downarrow)$
- Increasing the threshold β — defined as $(\beta \uparrow)$
- Decreasing α and increasing β simultaneously — defined as $(\alpha \downarrow \beta \uparrow)$

d) The utility functions: The outcomes of choosing a specific strategy are measured using a payoff function. The utility function is defined to reflect a player's potential performance gains or benefits from pursuing a specific strategy. As previously mentioned, different threshold values effect the two players A and G. Considering a certain strategy profile, say (s_m, s_n) leading to thresholds (α, β), the associated payoffs of the players are described as follows:

$$u_A(s_m, s_n) = Accuracy(\alpha, \beta), \tag{9}$$

where u_A is the payoff function of player A, and $Accuracy(\alpha, \beta)$ is defined in Equation 6, and

$$u_G(s_m, s_n) = Generality(\alpha, \beta), \tag{10}$$

where u_G is the payoff function of player G, and $Generality(\alpha, \beta)$ is defined in Equation 7.

For player A and player G, a value of 1 refers to a maximum utility while a value of 0 reflects a minimum payoff.

3) The trade-off between accuracy and generality:

a) Determining the Nash equilibrium: The game is viewed as a competition between the accuracy and generality measures of clustering. This is highlighted in Table II, where the table's rows refer to the strategies of player A and the columns refer to the strategies of player G. Each cell in Table II corresponds to a strategy profile, (s_m, s_n), where s_m represents player A's strategy and s_n represents player G's strategy. The goal of each player is to choose a strategy that configures the (α, β) thresholds in order to maximize her/his utility. $u_A(s_m, s_n)$ and $u_G(s_m, s_n)$ are the payoffs for players A and G, respectively, according to the strategy profile (s_m, s_n).

The logic in a game is that a player chooses a strategy with a larger payoff over other strategies with a lower payoff. For the two-player game under consideration, a strategy profile will be Nash equilibrium, with respect to the definition given in Equation 8, if,

$$Accuracy : \forall s_m \in S_A, u_A(s_m, s_n) \geq u_A(s_m^{'}, s_m), \tag{11}$$

where $(s_m^{'} \neq s_m)$, and

$$Generality : \forall s_n \in S_G, u_G(s_m, s_n) \geq u_G(s_m, s_n^{'}), \tag{12}$$

where $(s_n^{'} \neq s_n)$. This signifies that no player will gain from changing her/his strategy other than the strategy specified by the profile (s_m, s_n).

b) Determining the changes in the thresholds: Essentially, there are four ways for changing the thresholds (α, β) [29]:

1) A single player proposes to decrease the value of α — denoted as $(\alpha-)$;
2) Both of the two-game players propose to decrease the value of α — denoted as $(\alpha - -)$;
3) A single player proposes to increase the value of β — denoted as $(\beta+)$;
4) Both of the two-game players propose to increase the value of β — denoted as $(\beta + +)$;

These four ways can be used to associate threshold pairs with a certain strategy profile. For example, a strategy profile with (s_2, s_2) which is equal to $(\beta \uparrow, \beta \uparrow)$ is represented as $(\alpha, \beta + +)$, since player A and player G propose to increase the value of β.

4) The learning mechanism defining the values of the thresholds: A single game run has minimal utility in terms of finding appropriate values for the (α, β) thresholds. A learning process will emerge as a result of iteratively changing the thresholds with the goal of improving the payoffs for the players. In this regard, the learning rule or criterion is based on the relationship between threshold modification and the influence on the players' utility. This relationship is used to define the four variables $(\alpha-, \alpha - -, \beta+, \beta + +)$. This is accomplished through the use of an iterative game.

Let (α, β) be the initial thresholds for a particular iteration of an iterative game. As previously mentioned, the Nash equilibrium will be utilized to compute and decide the game solution as well as the associated thresholds; which will be denoted as (α', β'). The four variables $(\alpha-, \alpha--, \beta+, \beta++)$ are calculated based on a fixed percentage of either increasing or decreasing the strategies' values in every iteration. For example, if the initial values of (α, β) = (1, 0), the percentage of increasing and decreasing the strategies is equal to 5%, and a strategy profile with (s_1, s_2) which equals to $(\alpha \downarrow, \beta \uparrow)$

TABLE II
PAYOFF TABLE FOR THE GAME.

		Generality (G)		
		$s_1 = \alpha\downarrow$	$s_2 = \beta\uparrow$	$s_3 = \alpha\downarrow\beta\uparrow$
Accuracy (A)	$s_1 = \alpha\downarrow$	$u_A(s_1, s_1)$, $u_G(s_1, s_1)$	$u_A(s_1, s_2)$, $u_G(s_1, s_2)$	$u_A(s_1, s_3)$, $u_G(s_1, s_3)$
	$s_2 = \beta\uparrow$	$u_A(s_2, s_1)$, $u_G(s_2, s_1)$	$u_A(s_2, s_2)$, $u_G(s_2, s_2)$	$u_A(s_2, s_3)$, $u_G(s_2, s_3)$
	$s_3 = \alpha\downarrow\beta\uparrow$	$u_A(s_3, s_1)$, $u_G(s_3, s_1)$	$u_A(s_3, s_2)$, $u_G(s_3, s_2)$	$u_A(s_3, s_3)$, $u_G(s_3, s_3)$

is represented as ($\alpha-$, $\beta+$). The new values of (α, β) = $(0.95, 0.05)$. The process can be halted once a satisfactory level of performance has been attained.

IV. Experiment setup

In this section, we will present a comprehensive description of the experimental setup for the three-way clustering with the GTRS approach to cluster Corticosteroid sensitivity with missing values.

A. Considered cohort

The used APROCCHS cohort includes patients who received corticotherapy and placebo treatment. A total number of 1233 patients is maintained after selecting the most important features, applying data enrichment, labeling the data, and deleting the duplicates (1 duplicate raw was found in the data and was deleted). In our preliminary study, and based on a ranking strategy, we worked with only 10 risk factors, presented in Table III, among the 24 features. The initial APROCCHS dataset contains 26 instances having missing values (i.e., 2%) which will form the set M.

B. Experimental Plan, Tests, and Tools

Our experimental protocol is divided into three stages. The first stage focuses on simulating data with missing values that aims to answer the question of the performance of the algorithm when adding more missing values. The second stage is devoted to exploring the impact of a parameter of the algorithm. Specifically, we study the impact of changing the value K of the nearest neighbors component which is part of the evaluation function $e(c_k, o_i)$ (see Section II-A). Finally, in the third stage, various percentages of the strategies' initial values are considered to study the influence of these values on the obtained results. Below is an outline of the three stages:

- Experiment 1: We evaluated the performance of the three-way clustering approach by using in each experiment several percentages of the missing values. As a first investigation, the algorithm was tested on four different missing data versions. The rate of missing values randomly chosen in this regard is based on 5%, 10%, 15%, and 20%. This experiment will enable us to respond to the following research question (**RQ1**): How can the percentage of missing values affect the performance of the algorithm?
- Experiment 2: The aim of this experiment is to explore the k nearest neighbors used in calculating the evaluation function to investigate its impact on the results. For this purpose, we choose to work with k = 5 and k = 7. Conducting this experiment will lead us to answer the following question (**RQ2**): Does the k nearest neighbors has an impact on the clustering results?
- Experiment 3: We assessed the choice of the strategies' initial values percentages and their effect on the obtained results. In this experiment, the algorithm takes as input a different set of $\alpha-$, $\alpha--$, $\beta+$ and $\beta++$. In our case, we tried to decrease α and increase β by 7% having initial values of $\alpha-$ equals to 0.93, $\alpha--$ equals to 0.86, $\beta+$ equals to 0.07, and $\beta++$ equals to 0.14, and by 10% having initial values of $\alpha-$ equals to 0.90, $\alpha--$ equals to 0.80, $\beta+$ equals to 0.10, and $\beta++$ equals to 0.20. By carrying out this experiment, we will be able to respond to the following question (**RQ3**): can the percentage of increasing and decreasing initial values of α and β influence the results?

Although the number of iterations is not defined, a maximum number is given to prevent the algorithm from continuing in an endless loop if it does not converge. While setting a maximum iteration of 20, the algorithm often converged between 3 and 4 iterations, based on the APROCCHS cohort. As for the clustering part, the k-means algorithm was used with k=2.

V. Results and discussion

A. Experimental results of GTRS approach

The results obtained from different GTRS-based approach runs with the various percentages of missing values inputs are shown in Tables IV – VII. The tables present the following observations:

- From Table IV (and similarly to all other Tables V – VII), it can be observed that in most runs the algorithm converge in the third iteration. We can also see how the thresholds are altered across the game's several iterations and how this affects generality and accuracy. For the experiment with 5% missing values, the initial thresholds of (α, β) = (1, 0) are set before the game starts, resulting in an accuracy of 0.98 and a generality of 0.88. However, in the second iteration, the accuracy and generality are still the same, while the threshold α is decreased and β increased by 0.14. For the experiment with 10% missing values, the accuracy is stable while the generality increased from 84% to 93%. For 15% and 20% missing values, we can notice a slight decrease in the accuracy (for 15% missing values: from 1 to 98%, for 20% missing values: from 99% to 96%) with an increase in generality (for 15% missing values: from 87% to 94%, for 20% missing values: from 83% to 91%) –

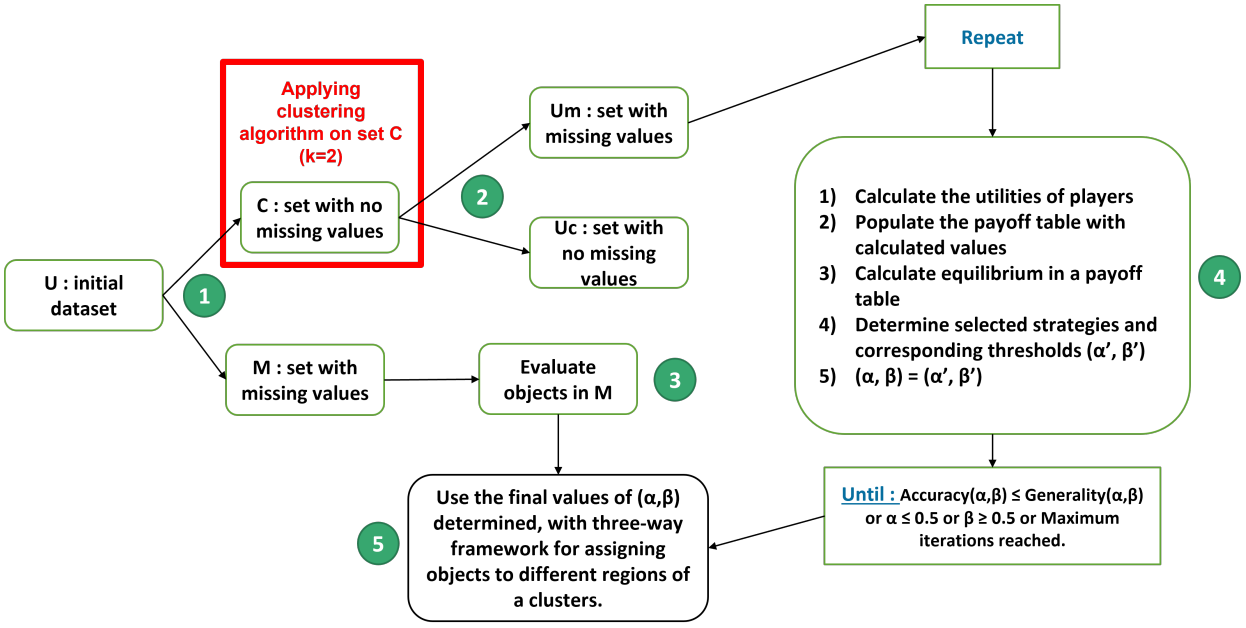

Fig. 1. Main functioning of the three-way clustering with Game-theoretic rough sets

TABLE III
CONSIDERED SET OF RELEVANT FEATURES AT DAY 0

Reference	Description	Format
DATINF	Diagnosis date	Precision = JJ/MM/YYYY, Min = DATHOSP (Hospital admission date), Max = Current date
SITINF	Infection location	0 = Lung, pleura, 1 = Peritoneal, 2 = Urogenital, 3 = Central Nervous System (CNS), 4 = Endocarditis, mediastinum, 5 = Sepsis, 6 = Soft tissue, 7 = Bones and joints, 8 = Other
SEX	Indicates patient sex	1 = Male, 2 = Female
PATWGHT	Indicates the weight of the patient	Min = 36, Max = 154
ORIGIN	Indicates the patient ORIGIN	1 = City, 2 = Hospital, 3 = Institution
AGE	Indicates patient age	Min = 18, Max = 97
KNAUS_J0	Activity and medical follow-up in the six months prior to admission	1 = Stage D Major activity restriction due to illness, including bedridden or hospitalized patients, 2 = Stage C Chronic illness causing significant but not total activity restriction, 3 = Stage B Moderate or moderate activity limitation due to illness (limited work activities), 4 = Stage A Good health, no activity limitation
MACCABE_J0	Description of the patient's condition before the episode leading to ICU	1 = Absence of underlying disease or underlying disease not life-threatening, 2 = Underlying disease life-threatening within 5 years, 3 = Underlying disease estimated to be fatal within one year
SOFA_ADM	Indicates the worst case value up to 3 hours after admission	Min = 2, Max = 16
IGSII_ADM_TYP	Indicates the admission type of the patient	0 = Scheduled surgery, 6 = Medical, 8 = Unscheduled surgery

a trade-off maintaining the required balance. To explore the research question **RQ1** of whether the percentage of missing values affects the performance of the algorithm, we performed the first experiment. By examining the outcomes of the GTRS algorithm via results presented in Tables IV – VII, when increasing the number of randomly chosen missing values, the trade-off accuracy/generality will not be lost.

- In the results presented in Table V, we have increased the value of k from 5 to 7. In comparison with the initial thresholds $(\alpha,\ \beta)$ = (1, 0) and when testing with only 5% of missing values, we can observe that there is a slight decrease in accuracy (i.e., 1%) while the generality improved with 9% reaching 95%. When testing with 10% and 20% missing values, the GTRS algorithm shows a minor reduction in accuracy varying from 1% to 3% with an increase in generality (between 8% and 13%). Thus, the GTRS model delivers an acceptable trade-off between accuracy and generality. For the experiment with 15%, and while comparing the results to the 5% missing values, we can note that there is an increase in the accuracy showing 98% (97% with 5% missing values) with a 1% decrease in generality. In order to investigate the research question of whether the value of

k nearest neighbors has an impact on the results **RQ2**, Experiment 2 was carried out. Through the interpretation of the results obtained from the GTRS algorithm, when increasing the percentage of missing values, and by increasing k to 7, we can notice a slighter loss in the trade-off accuracy/generality in comparison to k = 5.

- Table VI and Table VII show the results obtained when varying the strategies' initial values with decreasing α and increasing β by 10% having initial values of $\alpha-$ equals to 0.90, $\alpha--$ equals to 0.80, $\beta+$ equals to 0.10, and $\beta++$ equals to 0.20. Consistent with the findings in Table IV, from Table VI, we can notice that from 5% to 15% of missing values the accuracy demonstrates stability in its values with 98% while the generality presents a significant increase varying from 88% to 94%. For 20% missing values, the obtained results show a slight decrease in accuracy with 3% (reaching 96%) and a slight increase in generality with 2% (reaching 91%).
  As was the case in Table V, by analyzing the obtained results in Table VII, we can notice that when compared to the initial thresholds (α, β) = (1, 0) and tested with only 5% of missing values, we observed a slight decrease in accuracy (by 1%), but a significant improvement in generality (by 9%). Moreover, for experiments with 10%, 15%, and 20% one can observe that accuracy values were decreased by approximately 2% while generality increased by up to 14%. Also, for instance, with initial values equal to 7%, k = 5, and 20% of missing values (Table IV), the final values are 96% and 91% for accuracy and generality, respectively. With initial values equal to 10%, k = 5, and 20% of missing values (Table VI), the final values are the same registering 96% and 91% for accuracy and generality, respectively. To answer the third research question **RQ3** to what extent can variations in the percentage of initial values of α and β, whether increased or decreased, impact the results, Experiment 3 was implemented. By looking at the obtained results and interpreting them (Table IV, V, VI, and VII), it is noticeable that the final output of the GTRS algorithm is relatively stable regardless of the initial values of the strategies.

As expected, from the different tables, when using k = 5, the execution time is observed to be lower than when using k = 7, indicating that a smaller value of k can lead to faster computations. However, for more exploration, the execution time can be minimized by using several techniques such as Multi-threading [38], Single Instruction Multiple Data (SIMD) [39], and Open Multi-Processing (OpenMP) [40]. By employing these parallelism techniques, the GTRS algorithm execution time can be reduced, leading to marked improvements in both its performance and efficiency.

B. Three-way clustering approach evaluation

The previously obtained results show the effectiveness of the three-way clustering approach with GTRS in handling missing values. Therefore, in almost all the experiments, for clustering CS responsiveness, the trade-off accuracy/generality is maintained. The best trade-off found is with an accuracy value of 97%, and the generality presents 95%; with k =7 and 5% of missing values.

The final step (Figure 1 (3,5)) in the GTRS algorithm is to evaluate objects with missing values in M using Equation 4 and then select the best values of (α, β) and test them on the set M with missing values. As mentioned in Section IV-A, set M contains 26 patients having missing data (i,e., only 2%). Table VIII summarizes the obtained results after applying the three-way clustering approach on the set M and using k = 7 as value of k nearest neighbors. It can be observed that the accuracy/generality trade-off was preserved, presenting 96% accuracy and 92% generality. The results revealed that the best thresholds values for (α, β) = (0.58, 0.42). These final (α, β) values are used for assigning objects to different regions of a clusters as follows:

$$Inside(c_k) = \{o_i \in U | e(c_k, o_i) \geq 0.58\}, \tag{13}$$

$$Outside(c_k) = \{o_i \in U | e(ck, o_i) \leq 0.42\}, \tag{14}$$

$$Partial(c_k) = \{o_i \in U | 0.42 < e(c_k, o_i) < 0.58\}, \tag{15}$$

.

After applying Equations 13, 14, and 15 to assign objects in set M to clusters, we observed that the algorithm's non-deterministic nature resulted in some sepsis patients being found in the partial region. This means that these sepsis patients could not be clustered to a specific region as CS(placebo) sensitive(improved) or resistant(not improved); despite that we had their correct label in the cohort. In addition to this, when we examined the patients clustered by GTRS, we found some false negatives. This suggests that the results were not entirely deterministic, and further statistical analysis is required to validate them. One possible explanation to these preliminary results is that we have only considered 10 risk factors out of the 24 variables. Despite this, we still can consider that the initial results in terms of trade-off accuracy and generality are promising and indicate that GTRS has potential in addressing the issue of missing data in sepsis patients.

VI. Conclusion

The aim of this paper is to investigate the issue of clustering with missing values in clinical data using a three-way approach with GTRS. The study utilized data from the APPROCHS cohort, which included 1240 sepsis patients enrolled in a randomized controlled trial, and collected by clinicians from APHP. An important challenge in implementing this approach was setting appropriate thresholds to determine the three types of decisions. GTRS was found to be a promising alternative for clustering objects with missing values.

To evaluate the effectiveness of the GTRS model, three experiments were conducted. In the first experiment, the algorithm was tested with varying percentages of missing data, and the results showed that accuracy and generality can be preserved despite an increase in the number of missing

TABLE IV

OBTAINED RESULTS OF GTRS ALGORITHM APPLIED ON APROCCHS COHORT USING MULTIPLE MISSING VALUES PERCENTAGES AND FIXED VALUES OF K = 5 AND INITIAL VALUES = 7%

Missing values	k	Initial values	Iteration	Alpha	Beta	Accuracy	Generality	Execution Time
5%	5	7%	1	1	0	0.98	0.88	22 min
			2	**0.86**	**0,14**	**0.98**	**0.88**	
10%	5	7%	1	1	0	0.98	0.84	47 min
			2	0.86	0.14	0.98	0.84	
			3	**0.72**	**0.28**	**0.98**	**0.93**	
15%	5	7%	1	1	0	1	0.87	79 min
			2	0.86	0.14	1	0.87	
			3	**0.72**	**0.28**	**0.98**	**0.94**	
20 %	5	7%	1	1	0	0.99	0.83	154 min
			2	0.86	0.14	0.99	0.83	
			3	**0.72**	**0.28**	**0.96**	**0.91**	

TABLE V

OBTAINED RESULTS OF GTRS ALGORITHM APPLIED ON APROCCHS COHORT USING MULTIPLE MISSING VALUES PERCENTAGES AND FIXED VALUES OF K = 7 AND INITIAL VALUES = 7%

Missing values	k	Initial values	Iteration	Alpha	Beta	Accuracy	Generality	Execution Time
5%	7	7%	1	1	0	0.98	0.86	28 min
			2	0.93	0.07	0.98	0.86	
			3	0.72	0.28	0.98	0.89	
			4	**0.58**	**0.42**	**0.97**	**0.95**	
10%	7	7%	1	1	0	0.99	0.8	54 min
			2	0.86	0.14	0.99	0.8	
			3	0.72	0.28	0.96	0.88	
			4	**0.58**	**0.42**	**0.95**	**0.93**	
15%	7	7%	1	1	0	1	0.87	78 min
			2	0.86	0.14	1	0.87	
			3	**0.72**	**0.28**	**0.98**	**0.94**	
20%	7	7%	1	1	0	0.99	0.8	180 min
			2	0.86	0.14	0.99	0.8	
			3	0.72	0.28	0.99	0.85	
			4	**0.58**	**0.42**	**0.96**	**0.93**	

TABLE VI

OBTAINED RESULTS OF GTRS ALGORITHM APPLIED ON APROCCHS COHORT USING MULTIPLE MISSING VALUES PERCENTAGES AND FIXED VALUES OF K = 5 AND INITIAL VALUES = 10%

Missing values	k	Initial values	Iteration	Alpha	Beta	Accuracy	Generality	Execution Time
5%	5	10%	**1**	**1**	**0**	**0.98**	**0.88**	26 min
10%	5	10%	1	1	0	1	0.82	50 min
			2	0.8	0.2	0.98	0.9	
			3	**0.66**	**0.34**	**0.98**	**0.9**	
15%	5	10%	1	1	0	1	0.87	79 min
			2	0.8	0.2	0.98	0.94	
			3	**0.66**	**0.34**	**0.98**	**0.94**	
20 %	5	10%	1	1	0	0.99	0.82	104 min
			2	0.80	0.20	0.96	0.91	
			3	**0.66**	**0.34**	**0.96**	**0.91**	

TABLE VII

OBTAINED RESULTS OF GTRS ALGORITHM APPLIED ON APROCCHS COHORT USING MULTIPLE MISSING VALUES PERCENTAGES AND FIXED VALUES OF K = 7 AND INITIAL VALUES = 10%

Missing values	k	Initial values	Iteration	Alpha	Beta	Accuracy	Generality	Execution Time
5%	7	10%	1	1	0	0.98	0.86	28 min
			2	0.9	0.1	0.98	0.86	
			3	**0.7**	**0.3**	**0.97**	**0.95**	
10%	7	10%	1	1	0	1	0.78	39 min
			2	0.8	0.2	1	0.86	
			3	**0.66**	**0.34**	**0.98**	**0.92**	
15%	7	10%	1	1	0	1	0.83	78 min
			2	0.8	0.2	1	0.89	
			3	**0.66**	**0.34**	**0.99**	**0.94**	
20%	7	10%	1	1	0	0.99	0.80	106 min
			2	0.8	0.2	0.99	0.85	
			3	**0.66**	**0.34**	**0.96**	**0.93**	

TABLE VIII
BEST (α, β) VALUES EVALUATION ON THE SET M WITH MISSING VALUES

Missing values	k	Iteration	Alpha	Beta	Accuracy	Generality
2%	7	1	1	0	0.95	0.85
		2	0.86	0.14	0.95	0.85
		3	0.72	0.28	0.95	0.85
		4	**0.58**	**0.42**	**0.96**	**0.92**

values. The second experiment examined how the selection of the k nearest neighbors in the evaluation function affected the results. The third experiment evaluated the impact of the percentages of initial values of the strategies on the results, and the stability of the final output of the GTRS algorithm was apparent as it did not significantly vary with the initial values of the strategies.

As future work, we aim to use four clusters, instead of two, to further represent sepsis patients (Cortico-sensitive, Cortico-resistant, improved status with placebo, and unimproved status with placebo). This may improve the performance of the algorithm. Also, we aim to explore alternative approaches such as Reinforcement learning [41]. This approach would consider accuracy and generality as agents, and increasing and decreasing α and β strategies as actions to be taken in the environment. Players would learn a policy through trial and error that maximizes their rewards. Additionally, one can expand the evaluation of the results achieved by taking into account the quality of the model to address concerns related to overlearning [42], overfitting [43], and the assessment parameters used to measure the model's performance.

STATEMENTS OF ETHICAL APPROVAL

For the APROCCHS study, the protocol and qualification of all investigators were approved by the Ethics Committee (Comité de Protection des Personnes, CPP) of Saint-Germain-en-Laye, France, on November 22, 2007. The trial was registered at ClinicalTrials.gov under NCT00625209.

ACKNOWLEDGMENT

The national program "*Programme d'Investissements d'Avenir (PIA)*" (as part of the France 2030 programme) under the reference ANR-18-RHUS-0004. This work is part of the Federation Hospitalo-Universitaire (FHU) Saclay and Paris Seine Nord Endeavour to PerSonalize Interventions for Sepsis (SEPSIS). This work was also supported by ANR PIA funding: ANR-20-IDEES-0002 and by the iRECORDS project, funded by ERA PerMed (JTC_2021) to KZ and DA (ANR-21-PERM-0005)".

REFERENCES

[1] M. Singer, C. S. Deutschman, C. W. Seymour, M. Shankar-Hari, D. Annane, M. Bauer, R. Bellomo, G. R. Bernard, J.-D. Chiche, C. M. Coopersmith *et al.*, "The third international consensus definitions for sepsis and septic shock (sepsis-3)," *Jama*, vol. 315, no. 8, pp. 801–810, 2016.

[2] J. Matsuda, S. Kato, H. Yano, G. Nitta, T. Kono, T. Ikenouchi, K. Murata, M. Kanoh, Y. Inamura, T. Takamiya *et al.*, "The sequential organ failure assessment (sofa) score predicts mortality and neurological outcome in patients with post-cardiac arrest syndrome," *Journal of cardiology*, vol. 76, no. 3, pp. 295–302, 2020.

[3] K. E. Rudd, S. C. Johnson, K. M. Agesa, K. A. Shackelford, D. Tsoi, D. R. Kievlan, D. V. Colombara, K. S. Ikuta, N. Kissoon, S. Finfer *et al.*, "Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the global burden of disease study," *The Lancet*, vol. 395, no. 10219, pp. 200–211, 2020.

[4] A. Rhodes, L. E. Evans, W. Alhazzani, M. M. Levy, M. Antonelli, R. Ferrer, A. Kumar, J. E. Sevransky, C. L. Sprung, M. E. Nunnally *et al.*, "Surviving sepsis campaign: international guidelines for management of sepsis and septic shock: 2016," *Intensive care medicine*, vol. 43, no. 3, pp. 304–377, 2017.

[5] D. W. Cain and J. A. Cidlowski, "Immune regulation by glucocorticoids," *Nature Reviews Immunology*, vol. 17, no. 4, pp. 233–247, 2017.

[6] D. Annane, S. M. Pastores, W. Arlt, R. A. Balk, A. Beishuizen, J. Briegel, J. Carcillo, M. Christ-Crain, M. S. Cooper, P. E. Marik *et al.*, "Critical illness-related corticosteroid insufficiency (circi): a narrative review from a multispecialty task force of the society of critical care medicine (sccm) and the european society of intensive care medicine (esicm)," *Intensive care medicine*, vol. 43, no. 12, pp. 1781–1792, 2017.

[7] N. Heming, S. Sivanandamoorthy, P. Meng, R. Bounab, and D. Annane, "Immune effects of corticosteroids in sepsis," *Frontiers in Immunology*, p. 1736, 2018.

[8] D. Annane, "Corticosteroids for severe sepsis: an evidence-based guide for physicians," *Annals of intensive care*, vol. 1, no. 1, pp. 1–7, 2011.

[9] J. Cleve and U. Lämmel, *Data mining*. Walter de Gruyter GmbH & Co KG, 2020.

[10] S. Gavankar and S. Sawarkar, "Decision tree: Review of techniques for missing values at training, testing and compatibility," in *2015 3rd international conference on artificial intelligence, modelling and simulation (AIMS)*. IEEE, 2015, pp. 122–126.

[11] S. A. Alasadi and W. S. Bhaya, "Review of data preprocessing techniques in data mining," *Journal of Engineering and Applied Sciences*, vol. 12, no. 16, pp. 4102–4107, 2017.

[12] L. Yu, R. Zhou, R. Chen, and K. K. Lai, "Missing data preprocessing in credit classification: One-hot encoding or imputation?" *Emerging Markets Finance and Trade*, vol. 58, no. 2, pp. 472–482, 2022.

[13] M. S. Osman, A. M. Abu-Mahfouz, and P. R. Page, "A survey on data imputation techniques: Water distribution system as a use case," *IEEE Access*, vol. 6, pp. 63 279–63 291, 2018.

[14] R. R. Andridge and R. J. Little, "A review of hot deck imputation for survey non-response," *International statistical review*, vol. 78, no. 1, pp. 40–64, 2010.

[15] U. Pujianto, A. P. Wibawa, M. I. Akbar *et al.*, "K-nearest neighbor (k-nn) based missing data imputation," in *2019 5th International Conference on Science in Information Technology (ICSITech)*. IEEE, 2019, pp. 83–88.

[16] N. Karmitsa, S. Taheri, A. Bagirov, and P. Mäkinen, "Missing value imputation via clusterwise linear regression," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 4, pp. 1889–1901, 2020.

[17] R. A. Hughes, J. Heron, J. A. Sterne, and K. Tilling, "Accounting for missing data in statistical analyses: multiple imputation is not always the answer," *International journal of epidemiology*, vol. 48, no. 4, pp. 1294–1304, 2019.

[18] S. Goel and M. Tushir, "Different approaches for missing data handling in fuzzy clustering: a review," *Recent Advances in Electrical & Electronic Engineering (Formerly Recent Patents on Electrical & Electronic Engineering)*, vol. 13, no. 6, pp. 833–846, 2020.

[19] D.-T. Dinh, V.-N. Huynh, and S. Sriboonchitta, "Clustering mixed numerical and categorical data with missing values," *Information Sciences*, vol. 571, pp. 418–442, 2021.

[20] Z. Pawlak, "Rough sets," *International journal of computer & information sciences*, vol. 11, pp. 341–356, 1982.

[21] A. Skowron and D. Ślezak, "Rough sets turn 40: From information

systems to intelligent systems," in *2022 17th Conference on Computer Science and Intelligence Systems (FedCSIS)*. IEEE, 2022, pp. 23–34.
[22] T.-F. Fan, C.-J. Liau, and D.-R. Liu, "Variable precision fuzzy rough set based on relative cardinality," in *2012 Federated Conference on Computer Science and Information Systems (FedCSIS)*. IEEE, 2012, pp. 43–47.
[23] L. S. Riza, A. Janusz, C. Bergmeir, C. Cornelis, F. Herrera, D. Śle, J. M. Benítez *et al.*, "Implementing algorithms of rough set theory and fuzzy rough set theory in the r package "roughsets"," *Information sciences*, vol. 287, pp. 68–89, 2014.
[24] R. Polikar, J. DePasquale, H. S. Mohammed, G. Brown, and L. I. Kuncheva, "Learn++. mf: A random subspace approach for the missing feature problem," *Pattern Recognition*, vol. 43, no. 11, pp. 3817–3832, 2010.
[25] B. Panda, S. Gantayat, and A. Misra, "Rough set rule-based technique for the retrieval of missing data in malaria diseases diagnosis," *Computational Intelligence in Medical Informatics*, pp. 59–71, 2015.
[26] P. Maji, "Advances in rough set based hybrid approaches for medical image analysis," in *Rough Sets: International Joint Conference, IJCRS 2017, Olsztyn, Poland, July 3–7, 2017, Proceedings, Part I*. Springer, 2017, pp. 25–33.
[27] K. B. Nahato, K. N. Harichandran, K. Arputharaj *et al.*, "Knowledge mining from clinical datasets using rough sets and backpropagation neural network," *Computational and mathematical methods in medicine*, vol. 2015, 2015.
[28] Y. Yao *et al.*, "An outline of a theory of three-way decisions." in *RSCTC*, vol. 7413, 2012, pp. 1–17.
[29] M. K. Afridi, N. Azam, J. Yao, and E. Alanazi, "A three-way clustering approach for handling missing data using gtrs," *International Journal of Approximate Reasoning*, vol. 98, pp. 11–24, 2018.
[30] C. M. Poteraş and M. L. Mocanu, "Evaluation of an optimized k-means algorithm based on real data," in *2016 Federated Conference on Computer Science and Information Systems (FedCSIS)*. IEEE, 2016, pp. 831–835.
[31] J. P. Herbert and J. Yao, "Game-theoretic rough sets," *Fundamenta Informaticae*, vol. 108, no. 3-4, pp. 267–286, 2011.
[32] J. Yao and J. P. Herbert, "A game-theoretic perspective on rough set analysis," *Journal of Chongqing University of Posts and Telecommunications (Natural Science Edition)*, vol. 20, no. 3, pp. 291–298, 2008.
[33] N. Azam and J. Yao, "Analyzing uncertainties of probabilistic rough set regions with game-theoretic rough sets," *International journal of approximate reasoning*, vol. 55, no. 1, pp. 142–155, 2014.
[34] Y. Shoham, "Computer science and game theory," *Communications of the ACM*, vol. 51, no. 8, pp. 74–79, 2008.
[35] K. Leyton-Brown and Y. Shoham, *Essentials of game theory: A concise multidisciplinary introduction*. Springer Nature, 2022.
[36] R. S. Hotchkiss, E. Colston, S. Yende, D. C. Angus, L. L. Moldawer, E. D. Crouser, G. S. Martin, C. M. Coopersmith, S. Brakenridge, F. B. Mayr *et al.*, "Immune checkpoint inhibition in sepsis: a phase 1b randomized, placebo-controlled, single ascending dose study of anti-pd-l1 (bms-936559)," *Critical care medicine*, vol. 47, no. 5, p. 632, 2019.
[37] T. Z. J. Teng, J. K. T. Tan, S. Baey, S. K. Gunasekaran, S. P. Junnarkar, J. K. Low, C. W. T. Huey, and V. G. Shelat, "Sequential organ failure assessment score is superior to other prognostic indices in acute pancreatitis," *World Journal of Critical Care Medicine*, vol. 10, no. 6, p. 355, 2021.
[38] I. Oz and S. Arslan, "A survey on multithreading alternatives for soft error fault tolerance," *ACM Computing Surveys (CSUR)*, vol. 52, no. 2, pp. 1–38, 2019.
[39] M. Steinegger, M. Meier, M. Mirdita, H. Vöhringer, S. J. Haunsberger, and J. Söding, "Hh-suite3 for fast remote homology detection and deep protein annotation," *BMC bioinformatics*, vol. 20, no. 1, pp. 1–15, 2019.
[40] S. Bernabé, C. García, R. Fernández-Beltran, M. E. Paoletti, J. M. Haut, J. Plaza, and A. Plaza, "Open multi-processing acceleration for unsupervised land cover categorization using probabilistic latent semantic analysis," in *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2019, pp. 9835–9838.
[41] Y. Li, E. Fadda, D. Manerba, R. Tadei, and O. Terzo, "Reinforcement learning algorithms for online single-machine scheduling," in *2020 15th Conference on Computer Science and Information Systems (FedCSIS)*. IEEE, 2020, pp. 277–283.
[42] C. Song and V. Shmatikov, "Overlearning reveals sensitive attributes," *arXiv preprint arXiv:1905.11742*, 2019.
[43] X. Ying, "An overview of overfitting and its solutions," in *Journal of physics: Conference series*, vol. 1168. IOP Publishing, 2019, p. 022022.