Healthcare Analytics 1 (2021) 100001

Contents lists available at ScienceDirect

Healthcare Analytics

journal homepage: www.elsevier.com/locate/health

Features and explainable methods for cytokines analysis of Dry Eye Disease in HIV infected patients

Francesco Curia

Deparment of Statistical Science, Sapienza University of Rome, piazzale Aldo Moro 5, 00185, Rome, Italy

ARTICLE INFO

Keywords:
Clustering ensemble
Dry eye disease
Explainable artificial intelligence
Machine learning
Interpretable models
Features importance

ABSTRACT

Clinical Decision Support Systems (CDSS) that use machine learning techniques and their broadest sense of artificial intelligence (AI) must be interpretable and transparent. The lack of transparency instead of providing support could instead become a factor of indecision and obstacle. In this work, a very complex and important problem from a clinical point of view is tackled, namely the pathology known as Dry Eye Disease (DED), starting from a case-control study on a HIV-positive population and a healthy part of it. The case study is faced on two fronts, the first in which an ensemble-based clustering algorithm is built. Secondly, this algorithm is broken down to analyze each component, making the analysis method transparent and interpretable. Specifically, an ensemble of clustering algorithms is presented, such as k-means, agglomerative, spectral, and birch, which are combined and used in two levels: in the first, the labels are obtained from each clusterizer to recognize significant patterns of the two populations affected by the DED pathology, in the presence of HIV and not. Subsequently, the labels obtained at the first level are used as inputs on which the clusterizers are used again, whose outputs in the final phase serve as a training data set for a supervised method (i.e., logistic regression, decision trees, neural network, etc.), to evaluate every single component separately, through the use of features importance techniques (i.e., decision trees, LASSO regression, Gini Importance (GI), Variable Importance (VI), etc.). In this way, each clustering algorithm used at the first level can be considered a new feature in the next one and evaluate its individual contribution. Furthermore, each characteristic is interpreted through specific methods of the relevance of the characteristics to make the decision support tool as complete as possible. The performance of the methods used in training, both supervised and unsupervised, is evaluated through appropriate metrics, such as the well-known measures of precision, recall, accuracy, and homogeneity. Clustering methods provide results on the groups created and on the influence of features (cytokines) in the two populations examined. The experimental results obtained concerning the association between the development of the DED pathology and the presence or absence of HIV in these patients, and the influence that certain factors have on this problem, are interpreted with methods that are part of that branch known as Explainable AI (i.e., Local Interpretable Model-agnostic Explanations (LIME), Shapley, Individual Conditional Expectation (ICE), etc.). Besides explaining the influence exerted by certain features, the methods used provide both a global and local view on how each factor influences the final probability associated with the possible development of the pathology. The practical implications in using this method can be of support to the clinical diagnoses carried out on the patients examined to evaluate how each factor can be responsible for the possible development of the disease and therefore taken individually in the treatment. To date, the analytical techniques used in the study of this pathology have always provided generalized results, while breaking down the problem and isolating the components could provide valuable information to clinical operators.

1. Introduction

1.1. Clinical Decision Support Systems

Clinical Decision Support Systems (CDSS) play a important role in the clinical sector since every action taken by a decision maker is crucial from an ethical and legal point of view. Decision makers can be of various types in a CDSS picture, for example a doctor of medicine (M.D.), a minister or a task force of scientists (the current COVID 19 pandemic is a recent example of how political and clinical decision makers are called together to make important decisions). The results of a decision deriving from a decision support system that makes use of machine learning on the one hand make it possible make conscious choices, since it is assumed to be there an expert who oversees the decision-making process and on the other hand they

E-mail address: francesco.curia@uniroma1.it.

https://doi.org/10.1016/j.health.2021.100001
Received 30 April 2021; Received in revised form 2 July 2021; Accepted 2 July 2021

can lead to interpretable results, depending on which models They are used. A CDSS based on black-box (i.e. neural network) methods carries with them a great responsibility. The output of a model it may concern, for example, a drug therapy, administration of a drug, rather the experimentation of a vaccine that compatibility on organ transplants. The interpretability and transparency of the models used must guarantee full explainability of the results. Some questions are important, such as: why was one model preferred over another? and how this model was used. From the 1970s onwards there have been several CDSSs based on artificial intelligence, for example a work by de Dombal et al. ([1], 1972) in which you try to implement automatic reasoning in conditions of uncertainty. The system was developed from the University of Leeds, designed to support diagnosis of acute abdominal pain and on the basis of the analysis the need for surgery, system decision making it was based on the Bayesian approach. In Shortliffe's work ([2], 1976), (MYCIN), a rules-based expert system designed to diagnose e recommend treatment for certain blood infections (antimicrobial selection for patients with bacteremia o meningitis) has been proposed. It was later extended to management other infectious diseases. Clinical knowledge in CDSS is represented as a set of IF-THEN rules. Some CDSS related issues will be presented below for certain classes of problems, such as cancer, diabetes, heart problems and other applications; the use of advanced analysis techniques related to clinical decisions is an important topic and the literature is very broad, the most interesting contributions will be highlighted. Miller et al. ([3], 1982) developed INTERNIST-1, one of the first clinical decisions support systems designed to support diagnosis, in 1970. The CDSS was a rules-based expert system designed from the University of Pittsburgh in 1974 for the diagnosis of complex diagnoses of complex problems in general internal medicine. Use patient observations deduce a list of compatible disease states (based on a tree structured database that links diseases with symptoms). In the work of Muller et al. ([4], 2020) by definition these systems are based on patient specificity as evidence and representations of clinical knowledge modeled by algorithms and mathematical models by experts and provide recommendations by addressing the right diagnosis or optimal therapy. In this work the authors propose an approach based on data visualization. The authors show that more the displays show the certainty of the calculation result such as the recommendation and a series of clinical scores. Regarding the model used, the authors presented an approach for a CDSS based on a Bayesian causal network represents the therapy of laryngeal carcinoma. The results were evaluated and validated by two experts otolaryngologists. Several other studies have addressed the question of the explainability of CDSS, as in ([5], 2017), ([6,7], 2019), not calibrating the user trust concept by introducing this new type of error to the context analyzed by ([8], 2020) using these tools. Another example related to Bussone's work et al. ([9], 2015) who studied the effect of the explanation on trust and dependence. The authors state: "neglecting human factors and user experience in the design the explanation of the CDSS could lead to excessive dependence on medical professionals in these referral systems, even when it is wrong", which the authors define an "over-reliance". There is also another possible problem when the explanation it does not provide sufficient information could lead users to reject the suggestions, for example self-sufficiency as described in the work of ([10], 2020). There are other very recent works dealing with CDSS that make use of advanced techniques analytics, as in the work of ([11], 2020) in which a longitudinal retrospective observational study is conducted that examines 34,113 electronic medical records. The authors however use a multivariate logistic regression e time series analysis in order to explore the effects of CDSS. The aim of the study is to evaluate the effects of CDSS integrated with the British Medical Journal (BMJ) Best Practice Assisted Diagnosis in real-world research. With regard to the results they obtain total accuracy values of the diagnosis recommended by CDSS equal to 75.46% in the first degree diagnosis, and 83.94% in the top-2 diagnosis while 87.53% in the top-3 diagnosis in the data before implementation of the CDSS. The proportion of hospitalization time 7 days or less increased significantly by 7.83% (95% CI 1.79%–13.87%, P = 0.01). The authors therefore conclude that CDSS integrated with BMJ Best Practice has improved the accuracy of doctors' diagnoses. In another fairly recent study, the authors ([12], 2020) carry out an examination of 60 clinical support systems that use machine learning and find use in different clinical areas such as bacterial infections, viral infections, tuberculosis and on generic infections. 33% of these studies dealt with the diagnosis while 30% with the prediction of diseases, the prediction of the response to treatment and the prediction of antibiotic resistance, rather than the choice of antibiotic therapy itself. Regarding the implications, the authors themselves suggest that a data base as exhaustive as possible that takes into account factors such as primary care and socio-economic data can help to build much more effective tools. Other authors ([13], 2020) have addressed the study and prediction of heart disease with important results. Using spatial clustering techniques based on the density of applications with noise (DBSCAN) able to identify anomalies and remove them and then use a technique known as (SMOTE-ENN) to balance the distribution of train data and subsequently train an XGboost to predict heart disease. The authors compare their results with others already known in the literature, obtaining accuracies of 95.90% and 98.40% respectively, thus providing a tool that can be fully used by clinical operators. Recent work ([14], 2021) describes the prevalence and nature of the involvement of clinical experts in the development, evaluation and implementation of CDSS that use machine learning to analyze electronic health record data. The authors conduct a systematic search on different platforms such as: PubMed, CINAHL and IEEE Xplore and a manual search of conference proceedings in order to identify suitable articles. The results they get are quite interesting: the involvement of clinical experts was prevalent in the early and late stages of system design. The authors pay attention to the fact that clinical operators must necessarily be involved in the entire decision-making process in order to obtain a robust tool, in which therefore the clinical domain competence supports the analytical design phase designed by an expert, but which falls outside the medical domain.

1.2. Dry Eye Disease problem

DED (Dry Eye Disease), it is a condition of the human eye which occurs when the tears necessary for adequate lubrication for the eyes, occur in scarce quantities or almost absent, creating a disabling tear instability. This problem affecting the external sense organ of the visual apparatus leads to inflammation and possible damage to the surface of the eye. According to the American Optometric Association (www.aoa.org) DED can develop for many reasons, including:

- **Age** Dry eyes are part of the natural aging process. People over the age of 65 experience some symptoms of dry eye.
- **Gender** Women are more likely to increase dry eyes due to hormonal changes caused by pregnancy, oral contraceptives and menopause.
- **Medications** Some medications, including antihistamines, decongestants, blood pressure medications, and antidepressants, can reduce tear production.
- **Medical conditions** People with rheumatoid arthritis, diabetes, and thyroid problems are more likely to have dry eye symptoms. Environmental conditions. Exposure to smoke, wind, and dry climates can increase tear evaporation with symptoms of dry eye. Also the inability to blink regularly, such as when staring at a computer screen for long periods of drying the eyes.

As part of the CDSS, some studies have been conducted on this pathological condition; the authors ([15], 2019) starting from the factors that characterize the disease, such as those listed above and according to the guidelines of the American Academy of Ophthalmology, acquire various data concerning the disease in order to build a robust model to support clinical decision making to try to predict the

condition in advance by analyzing symptoms. The authors consider models such as neural networks, decision trees, random forest and naive Bayes. The results that the authors obtained were quite accurate, starting with the classification by decision trees given a sufficient amount of data, structured in a certain way. The prediction rate of random forest and decision tree algorithms is over 90% compared to more complex methods such as neural networks and naive Bayes. A more recent study ([16], 2020) developed a model based on machine learning methods such as decision tree and LASSO to then predict a scoring (probability) score for a classification using multiple logistic regression. The authors consider as many factors as possible from the data provided by the Korea National Health and Nutrition Examination Survey (KNHANES) for 2012 (4391 sample cases). The results obtained show that the point-based model obtained an AUC (area under curve) of 0.70 (95% CI 0.61–0.78). Important factors included gender (+9 points for women), corneal refractive surgery (+9 points), current depression (+7 points), cataract surgery (+7 points), stress (+6 points), age (54–66 years; +4 points), rhinitis (+4 points), lipid-lowering drug (+4 points) and omega-3 intake (0.43%–0.65% kcal/day; −4 points). The proposed method is valid for finding important risk factors and identifying the patient's specific risk that could be applied to other multifactorial diseases.

1.3. Objectives

As seen in Sections 1.1–1.2, CDSS can provide advanced tools in the fight against various diseases, making use of advanced machine learning and deep learning techniques; the purpose of this work and objectives can be spelled out below:

- Addressing a complex problem such as DED disease related to HIV status in HIV infected patients, in order to make a comparison with a healthy population and try to infer characteristics that may be of interest in studying the development of the disease
- Show that through advanced methods of machine learning, both supervised and unsupervised, it is possible to direct research in this area towards more recent technologies; studies so far at the medical health level, make use of classical statistical tools which, however important, have intrinsically distributive hypotheses that cannot always be satisfied
- Analyze the factors that influence the development of the disease through methods that can be interpretable and constitute an advanced means of diagnostics, analyzing both locally each single factor and as a whole, in order to have a broader picture of the disease

1.4. Implications

In this work there are several implications that can bring added value to the study of CDSS both in the specific context of this DED disease, and for other problems that can be addressed by this approach. First, a combination of unsupervised and supervised method is carried out. The use of clustering techniques provides evidence on the data structure, patterns and elements that can be extracted for information and diagnostic purposes. Often, however, these techniques are an end in themselves, in the sense that once groups have been created, the evidence and correlations between the factors and elements that make up the clusters are sought. In this paper, however, the clustering methods are used twice simultaneously and using the output of the first training cycle allows you to use it as input in the second and get an overview of which method is better. Once this is done, it is subsequently possible to predict or classify an instance by using a meta-regressor or meta-classifier, thus being able to study the probability of assignment to a particular cluster and evaluate the influence of each individual feature. The approach is totally new, as for the works cited previously, both for the classic CDSS and for those inherent to DED, none of the cited authors has provided explanations and interpretability of both the model used and the results obtained, thus providing the clinical decision maker an instrument that it can be defined as "blind". Black-boxing methods are certainly reliable and accurate, but they must provide answers to the decision maker. The results obtained are of great interest, as in addition to being able to establish whether a particular patient may belong to one group rather than another, the method provides results in terms of probability of disease development and the possibility of opening the model and individually evaluate each method used (for clustering) and the relationship of the features (for supervised classification).

1.5. Outline

As regards the structure of the work, it is divided as follows: an introductory part in which the panorama of CDSS methods is presented and a general framework on Dry Eye Disease, with state of the art and main case studies, implications, technologies and limitations. In the introduction, the objectives and implications of this work are presented. Below is a part related to the method presented in this work (Related work): desirable properties of interpretable CDSS, the reasons for using this approach for DED, a description of the data, the mathematical methodology with the description of a clustering algorithm based on stacking method. Subsequently a part on explainability (Explainable ML) is presented with the main methods of features explanation and feature importance, both for supervised and unsupervised methods. The fourth part of the work (Experimental results) concerns the results obtained in terms of performance of the algorithms used, the explanation (opening black-boxe) of the algorithms. The fifth part (Clinical Explainability) discusses the part relating to supervised and unsupervised methods but from a clinical point of view, on the relationships and implications of the different factors that make up the analysis of the DED disease. In the last part the conclusions follow with a brief summary of what has been done and what has been discussed, to then address the limitations of the work and future objectives.

2. Related work

A recent work [17] compared two populations (HIV positive, n = 17 and healthy controls, n = 18) in order to assess whether there was an association between the pathology and dropout of the meibomian gland; the authors found statistically significant associations in the group of HIV-infected individuals. This condition has been found in 50%–80% of cases in HIV and AIDS patients. A highly significant CD4 cell count has been associated with this condition, correlated with a serious situation of the eye, as indicated by a recent study [18]. Starting from the case study by Agrawal et al. concerning a case-control [19], [20] will be treated in patients with HIV infection (type 1). The authors review and compare data from 34 HIV-infected patients and 32 control patient observations, in order to: "study the profile of tear cytokines in HIV infected patients with HIV Disease Dry Eye (DED) and studies the association between the severity of ocular inflammatory complications and tear cytokine levels". The proposed methodology by the authors, however, it is not about a machine-driven study learning methodologies but rather a parametric study based on classical statistics epidemiological approach; in this application the goal is to find meaningful models by the grouping procedure of the whole. The method it is therefore unsupervised despite the presence (if desired) of a binary variable which indicates whether the patient is HIV infected or not. Factors that make dry eyes more likely may include the fact that tear production tends to decrease as the age. This condition generally occurs in female individuals over the age of 50, in many cases due to hormonal changes caused by pregnancy, or due to the use of the birth control pill or even due to a menopause issue. Diets low in vitamin A or low in omega-3 fatty acids can contribute to this condition, not least wearing contact lenses may be among the causes of the development of this pathology. It is not a discussion of this application to predict

whether a patient with certain characteristics may be affected by the disease although it is not excluded that it may be a topic for later discussion. The study [19] involved the comparison of 41 features inherent in cytokine levels using the Luminex bead assay [20]; the authors collected the data through recruitment in a Singapore referral eye center. The authors used logistic regression for the study in order to understand the correlations and the statistical significance of the relationships. As mentioned, the intent of this work is to find significant patterns in the data, through an unsupervised ensemble method in accordance with the results obtained by the authors, they state that specifically: "statistically significant differences were observed in the mean epithelial growth factor (EGF), growth-related oncogene (GRO) and gamma-induced interferon values protein 10 (IP-10)". They also state that" EGF and IP-10 levels were higher and GRO levels were lower in DED tear HIV-infected patients compared with DED patients without HIV infection. The authors found: "no significant association between varying levels of ocular surface parameters and cytokine concentrations in HIV patients with DED", for a p-value greater than 0.05. The authors therefore conclude that: "the EGF and IP-10 values were significantly elevated and the GRO levels were lower in the tear profile of HIV patients with DED versus immunocompetent patients with DED".

2.1. Motivation

Clinical Decision Support Systems (CDSS) have always played a very important role: from cancer problems [21], to heart attack prevention [22], rather than the study and treatment of diabetes [23], numerous CDSS tools use intelligent systems that rely on advanced machine learning and deep learning techniques [24], therefore the role played by these complex algorithms has become a leading role in recent years. Providing clinical operators with transparent tools is a responsibility that researchers in the field of AI must take on and therefore in recent years a new branch known as Explainable AI has been born, in order to make these so-called black-boxing, interpretable and transparent by the point of view of how these outputs are obtained by the algorithms, [25]. This work offers interpretable and transparent tools in order to build a CDSS that respects some desirable properties of a decision support system [26] that makes use of machine learning or deep learning techniques and methods. Such desirable properties are the simulability, decomposition and transparency of the algorithms (Fig. 3).

1. **Simulability**: The author [26] sets this sub-feature as follows: "a model is transparent if a person can contemplate a model simultaneously" or in the meaning of Ribeiro et al. [27]: "a model is interpretable if it can be presented visually and understood intuitively". The concept of transparency is also applicable as the algorithm training is provided, the necessary steps and each step which output it produces. Of course, by definition an interpretable model is a simple model, but a problematic problem that is added to that of interpretation is surely the concept of compromise between complexity and interpretability, the challenge therefore remains methods that work intuitively dare a fairly simple explanation of complex models, such as which on average lead to very predictive results more accurate.
2. **Decomposition**: Lipton [28], in his work he also introduces the decomposition property, i.e. each input element of a model must be individually interpretable, just think of the inputs of a classification tree or a linear model. This property discriminates between models to which the inputs are engineered or anonymous, this property could also be affected not only by the number of features considered and by their engineering, even if certain indicators are contained in the data set or not.
3. **Algorithmic transparency**: At the algorithm level the notion of interpretability can be applied in the phase in which the algorithm learns. The author sets the example in the case of linear models, investigating the surface of the error obtained by minimizing the loss function. The convergence to an excellent global also for test data, can introduce the concept of trust in the learning method, which in black-boxing like deep neural networks does not happen because, in the training phase, often the associated cost functions to learning, they are optimized through heuristic methods and therefore the solutions produced are not optimal overall, but at best with approximations. Therefore the concept of algorithmic transparency lies precisely in the very way in which the algorithm works.

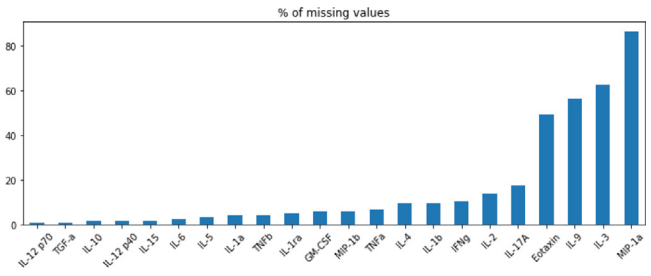

Fig. 1. The figure shows the percentage values of missing values for the features present in the dataset: over 10% it is advisable to remove such features that would not add value to the analysis.

The aim of this work is therefore to provide a clinical support tool based on advanced analytical methods that make use of intelligent systems, such as ensemble clustering and methods of explainability of algorithms and interpretation of results; the topic dealt with is complex and therefore it is important from a clinical point of view to be able, in addition to having an output on the risk of developing this pathology (DED), it is equally important to break down the problem and understand the relationships and weight of each factor inside the built system. The methods used and the results obtained from this study will be presented in the next sections.

2.2. Dataset

The data [20] concern 41 cytokine-related characteristics of HIV-infected patients with DED (n = 34) and unaffected patients (n = 32) for a total of 126 observations and 44 features (patient id, binary target that indicate if HIV it is present or not and binary feature indicate which eye is involved, right or left eye), these data were acquired through analyzes carried out at a clinical facility in Singapore. The data were processed by excluding the features that presented a percentage of missing values > 10% (Fig. 1). Therefore the following have been excluded: **Eotaxin** 49%, **IL-17A** 17%, **IL-2** 13% and **IL-3** 62%, **IL-9** 56% and **MIP-1a** 86%. The variables that had values lower than 10% were imputed through the mean of the variable. In spite of this data set, in order to implement a new unsupervised method, neither the binary target variable indicating the presence or absence of HIV and the binary variable indicating whether the eye is the right or the left is not considered; these two features will be used later in the visualization of the data and in the part relating to the features explainability.

2.3. Methodology

In most machine learning problems, the target variable that indicates the presence or absence of the phenomenon under study (in the case of classification, both binary and multiclass) is not always present; there are cases in which learning problems are also so-called unbalanced, when the proportion of cases in a classification problem is very different and leans more towards one class than another. Training an unsupervised clusterizer is generally an excellent method for inferring information within the data structure and for trying to understand if

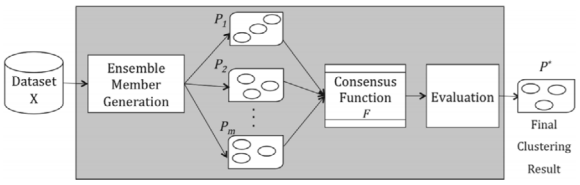

Fig. 2. A generic clustering ensemble framework, Source: Alqurashi, T. and Wang, W."Clustering ensemble method".

a certain method can actually be defined as intelligent. In order to understand the value and influence that cytokines have on the complex clinical picture of the DED disease, it was deliberately decided to omit the target variable and transform the problem into an unsupervised problem; in this way the analysis is totally addressed on the examined population and on the characteristics concerning the cytokines, in order to highlight if there are significant differences between the groups. To do this, a new method is therefore introduced, instead of considering and analyzing the available data in a classical way. Obtaining groups and omitting the target variable is a good way to evaluate post-analysis the value of the deductions obtained on the data starting from only the set of features inherent to the cytokyne. The method proposed in this work is based on a new clustering methodology, precisely defined Stacking Clustering Algorithms, in which a series of $C_1(x), \dots, C_n(x)$ clusterizers are applied to the initial dataset in order to produce assignments of the respective examples to a specific cluster identified a unique label. This label, for each clusterizer used in the previous step will be used to train other m-clusterizers in the second step. This method is known as stacking ensemble [29], in which this procedure is used and the final prediction is obtained by applying a meta-learner (in the supervised case) while in the unsupervised case as shown in Fig. 2, the different clustering algorithms chosen are combined by the consensus function, which in our case is based on majority voting, expressed by the following formula

$$C^* = \text{mode}(C_1(\mathbf{x}), \dots, C_k(\mathbf{x})) \tag{1}$$

Once the final label for each cluster has been obtained from the consensus function (1), a meta-learner is trained who takes as input the set of meta-features and the optimal label as a target, in order to obtain a probability of belonging to a specific cluster and obtain a features importance for each cluster model used, as using the meta-features space the input becomes the cluster algorithm used and therefore it is possible to obtain an importance ranking of each of them; the steps described are shown in the pseudo code presented below.

The strength of this method is intuitive; selecting a set of algorithms that contribute to forming a final clustering model, through the conversion into a classification model through the introduction of a meta-learner on the meta-features space, allows to obtain the importance of each of the clusterizers used, being able in this way to decompose the ensemble and make it interpretable.

3. Explainable machine learning

The methods of Explainable AI are much discussed today and are beginning to play a very important role in the science of decision making, as an intelligent system often based on black-box methods must necessarily be able to provide the decision maker with the possibility of know

(a) how the decision came about
(b) how this decision is to be interpreted

this must necessarily be contemplated in the context of clinical decision support systems. In order to build a transparent and interpretable clinical decision support system, some of the main explainable ML methods used are introduced.

Algorithm 1: Stacking Clustering Algorithm

Phase 1
input:
Features set $\mathbf{X}$
k-clusterizers, $C_i, i = 1, ..., k$
for i ← 1 to k **do**:
Train C_i on features set $\mathbf{X}$ and obtain the cluster's label
$l_i = C_i(\mathbf{X}), i = 1, ..., k$
Assign
$l_i, ..., l_k \leftarrow \tilde{\mathbf{X}}$
Phase 2
input:
New features set $\tilde{\mathbf{X}}$
k-clusterizers, $C_i, i = 1, ..., k$
for i ←1 to k **do**:
Train C_i on new features set $\tilde{\mathbf{X}}$ and obtain the cluster's label
$\tilde{l}_i = C_i(\tilde{\mathbf{X}}), i = 1, ..., k$
do:
Compute $C^* = \text{mode}(\tilde{l}_1, ..., \tilde{l}_k)$
Phase 3
input:
Data $D = (\tilde{\mathbf{X}}, C^*)$
Supervised learner $\mathcal{L}$
do:
Train $\mathcal{L}$ on D and get p_j

Output: Probabilities class $p_j = P\left(C_j = i | X_j\right)$

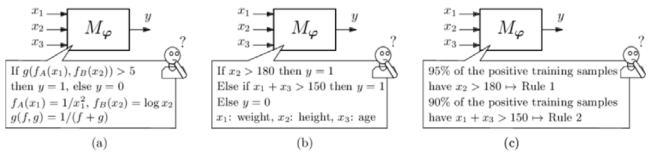

Fig. 3. Conceptual diagram exemplifying the different levels of transparency characterizing a ML model M with ϕ denoting the parameter set of the model at hand: (a) simulatability; (b) decomposability; (c) algorithmic transparency. Source: [26].

- **LIME**
  Ribeiro et al. [27] in this regard, introduces the concept of trade off between interpretability and loyalty LIME (Local Interpretable Model-Agnostic Explanations) formalized through the following optimization problem:

$$\min_{g \in G} L(f, g, \pi_x) + \Omega(g) \tag{2}$$

  where $\Omega(g)$ can be defined as a measure of complexity (as opposed to interpretability) of the model g, for example the number of parameters, or the depth of a tree in the case g is a Classification Trees, or for a linear model the number of non-zero weights, for example in the Lasso–Ridge approach. So a model g, belonging to the wider class of models G, minimizes the L, which is a loss function which measures the infidelity of the model considering the proximity measure π_x. Infidelity is defined by the authors as "the predictive behavior of the model near the instance to be predicted", therefore a discrepancy between what is expected and what is predicted.

- **Partial Dependence Plot**
  In Friedman's work [30] some methods for the interpretation of models are presented. PDP is focuses on visualization, one of the most powerful interpretative tools and the display is limited to small topics. Functions of a single variable with real value

can be plotted as a graph of the values of $\hat{F}(x)$ against each corresponding value of x. The functions of a single categorical variable can be represented by a bar chart, each bar represents one of its values and the bar height the value of the function. Viewing functions of higher-dimensional topics is more difficult. Is therefore useful to be able to visualize the partial dependence of the approximation $\hat{F}(x)$ on small selected subsets of the input variables. The functional form of $\hat{F}$ depends on the chosen values of the input subset z_l, if the dependency is not very strong the expected value of $\hat{F}(x)$, that is $\mathrm{E}[\hat{F}(x)]$ can represent a good synthesis of the partial dependence of the chosen variables of the subset z_l, a value such that $z_l \cup z_i = x$ where z_l is the complement subset of size l and z_i is a chosen target subset. Dependencies can be different, as additive or multiplicative, for example in classification problems the author suggests that partial dependence diagrams of each $\hat{F}_k(x)$ on subsets of variables z_l most relevant for a given class provide information on how input variables affect the respective probabilities of individual classes.

- **Individual Condition Expectation**
  ICE [31] is a tool to visualize the model estimated by any supervised learning algorithm. While the PDP helps to visualize the partial average relationship between the estimated response and one or more features, in the presence of substantial interaction effects, the partial response relationship can be heterogeneous, therefore an average like the PDP, can blur the complexity of the relationship modeled, instead the ICE improves the partial dependence diagram by graphically representing the functional relationship between the expected response and the characteristic for the individual observations. In particular, the ICE graphs show the variation of the values adapted in the range of a variable suggesting where and to what extent heterogeneity can exist.

- **Feature Interaction**
  Starting from his work on the PDP method, Friedman *et. al* presents another method, called Feature Interaction [32] which assumes that a function $F(x)$ has an interaction between two of its variables x_j and x_k if the difference in the value of $F(x)$ as a result of changing the value of x_j depends on the value of x_k. Such an assumption can be formalized as

$$\mathrm{E}_x\left(\frac{\partial^2 F(x)}{\partial x_j \partial x_k}\right)^2 > 0$$

  or by an analogous expression for categorical variables implying finite differences. If there is no interaction between these variables, the function $F(x)$ it can be expressed as the sum of two functions, that is $F(x) = f_j(x_j) + f_k(x_k)$ one of which does not depend on x_j and the other independent of x_k.

- **Shapley Value**
  Among the important works to refer to it is possible to mention the Shapley Values [33], an innovative method in which an additive method assesses the importance of variables through the expected conditional value of the original model, it is possible to mention the work of Koh and Liang [34] in which the authors measure the importance of the variables through the Influence Function, i.e. starting from the minimization of a risk function of the following type $R(\theta) = \frac{1}{n}\sum_i L(z_i, \theta)$. For a more detailed discussion, from which various components of this chapter have been extracted, please refer to the excellent work of the authors [26].

3.1. Features importance

In this part of the work are presented some of the main features importance methods in order to give the reader a general overview and understanding of the context in which this work is placed. Features importance is an analytical technique that aims to understand how much each feature within the data space contributes to the final prediction (or classification); the methods presented here are those widely known and applied in different contexts, including the clinical one, due to their simplicity of interpretation and explainability. The first two methods are a consequence of the application of the Random Forest algorithm, a set of predictors or classifiers of the decision tree type combined in a causal way in order to improve the final result of the model; these methods [35] are defined respectively variable importance (VI) and Gini importance (GI) which aim to evaluate the features in the model when it descends the impurity of the nodes at each iteration, permuting the features in a random way; the GI method evaluates this decrease through the Gini index, while the VI considers the average decrease. Another interesting method is the one proposed by Gedeon [36] who introduces a new method based on the matrix of the input weights of a neural network, through the random elimination of less important features using a brute force method. A recent and very interesting method [37] is the Importance Ranking Measure (FIRM), which uses the retrospective analysis of machine learning algorithms that allows to obtain both predictive performance and performance from the point of view of explainability. This method is also interesting as it considers the underlying correlation structure of the features in such a way as to find the most important features. Another interesting method is the PIMP [38] which is a heuristic correction of the VI and GI methods, in which the target variable is exchanged estimate the importance of a features in a casual way. assuming that it follows a certain probability distribution (Gaussian, lognormal or gamma), the value of the p-value obtained from the resulting estimate is used as a corrected measure of feature relevance.

3.2. Evaluation

The methodologies mentioned in the previous subsections 3 require some algorithmic performance evaluation measures; since the problem treated in this work involves a hybrid approach in which first unsupervised methods (clustering) and then a supervised meta-learner are applied in order to make the model interpretable and obtain a probability of belonging to a given cluster, it introduced both measures that allow to evaluate the homogeneity of the groupings obtained with stacking clustering and both measures that evaluate the correct classification of the instances; last but not least, the hybrid approach first unsupervised and then supervised makes it possible to apply the methods of explainability discussed above 3.

Clustering. Rosenberg and Julia Hirschberg [39] introduce a measure of homogeneity known as V-measure (Validity measure), based on the concept of external entropy that solves some problems that other evaluation metrics used in clustering present, such as the type of data processed, the algorithm used and also the simultaneous measurement of two desirable properties such as homogeneity and completeness. One of the traditionally used evaluation methods is the Dunn's index [40], which measures the internal homogeneity of the points grouped for each cluster, minimizing the internal variance and separating the groups externally. Another well-known and used method is the Silhouette Coefficient [41], which validates the measure of coherence within the clusters referring to the quality of the classification of each object; this method is widely used and provides an intuitive visual representation of the grouping.

Classification. Starting from the confusion matrix that it can be applied to binary and multiclass classification problems the algorithms used were evaluated by the following metrics, namely recall, precision, accuracy and Area Under Curve (AUC). The metrics defined are

$$precision = \frac{TP}{TP + FP} \tag{3}$$

and

$$recall = \frac{TP}{TP + FN} \tag{4}$$

For binary classification, accuracy can also be calculated in terms of positives and negatives classes

$$accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{5}$$

where

(a) TP (True positive): correctly classified or detected
(b) FP (False positive): incorrectly classified or detected
(c) FN (False negative): incorrectly rejected
(d) TN (True negative): correctly rejected

furthermore is defined the F-measure score as follow

$$F\text{-}measure = \frac{2}{1/precision + 1/recall} \tag{6}$$

for classification problems in which the probability of belonging to a certain class of an instance is evaluated, it is possible to use the following metric defined receiver operating characteristics (ROC), which is also a good intuitive visual interpretation, from this measure it is possible to derive the formulation of AUC which has a very interesting statistical property as shown by [42]: the AUC of a classifier is equivalent to the probability that the classifier will classify a randomly chosen positive instance higher than a randomly chosen negative instance.

4. Experimental results

This section presents the main results obtained, both for the ensemble clustering method and for the supervised part. The performance results of the models, the interpretation of the features, their importance and a whole part in which the results obtained from the clinical point of view are explained are presented.

4.1. Algorithms performance

For the part relating to clustering algorithms, four methods have been selected: k-means, agglomerative clustering, birch and spectral clustering. Using the silhouette method and the elbow method for determining the necessary clusters, three distinct clusters were created. The choice of these methods falls on their good interpretability, methods that answer well to the question "how the algorithm works" (Section 2.1): the k-means is based on a euclidean distance and from the point of view of mathematical optimization its resolution is very simple, it minimizes the variance between groups. The agglomerative algorithm is a hierarchical method in which in each step of the algorithm, the two most similar clusters are combined into a new larger cluster. This procedure is repeated until all points are members of one large cluster. The birch method is always part of the hierarchical methods and works a lot on large amounts of data, in our case study considering a large set of features it could be a good method, from the point of view of transparency, this method works through a tree structure therefore from the logical point of view it is easy to understand. Spectral clustering is a technique that works well where the set of features is very large as it reduces the dimensionality by mapping the new points in a vector subspace, making one of the adjacency matrix being a graph-based method; by choosing a euclidean kernel function the relationship with the k-means method is direct and therefore also its resolution. For the evaluation of the capacities of the single clustering algorithms it is possible to use the previously presented metric known as V-measure: it is evaluated the homogeneity and completeness between the optimal label obtained by the consensus function and the label of each single clusterizer, then it evaluates the homogeneity and completeness of each clusterizer with itself but at the previous level in staking, i.e. it evaluates the label obtained at the first level with that obtained at the second level for the same method. Then the results obtained are given in Table 1.

For the supervised part, the well-known logistic regression was used in order to create a meta classifier of the instances on the basis of the meta features previously obtained with the stacking clustering procedure. The classifier correctly assigns the respective classes to the classified test instances with good accuracy, as can be seen from Fig. 4 which shows the respective confusion matrices for the train and test set. The accuracy of the classifier for the train and test set data, respectively, is equal to 0.91 and 0.85 with AUC value 0.86, Table 2 shows the values of the other metrics obtained for each clustering label on which the classification was carried out.

Fig. 5 shows the analysis of the ROC curve with respect to the three analyzed clusters; the relationship between true positives and false positives for cluster 0 is quite good in correspondence, respectively, of values equal to 1 (TP) and 0.1 (FP), while for cluster 1 the situation is a little different, it is respectively 1 (TP) and 0.5 (FP), the last cluster, 2 instead shows values equal to 1 (TP) and over 0.5 for false positives (FP), being the AUC equal to 0.86 it can be deduced that there is 86% probability that the result of this classifier applied to an individual randomly extracted from the group of patients is higher than that obtained by applying it to an individual randomly extracted from the group of healthy, therefore a good chance of being able to distinguish in which cluster to assign the new instances.

Table 1
V-measure analysis.

Method	Algos-labels and optimal label	Intra-stacking levels
birch	0.73	1.0
agglomerative	1.0	0.67
k-means	1.0	0.40
spectral	0.73	0.38

Table 2
Classification report.

	Precision	Recall	F1-score
cluster 0	0.90	0.90	0.90
cluster 1	0.82	0.90	0.86
cluster 2	0.88	0.78	0.82

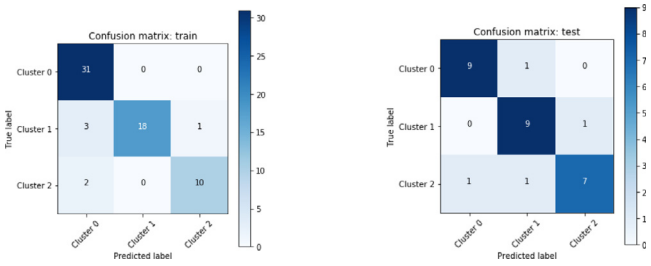

Fig. 4. Confusion matrix comparison: train vs test.

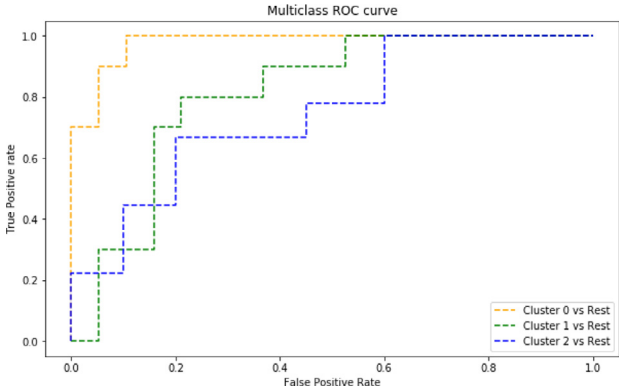

Fig. 5. Multiclass ROC curve analysis.

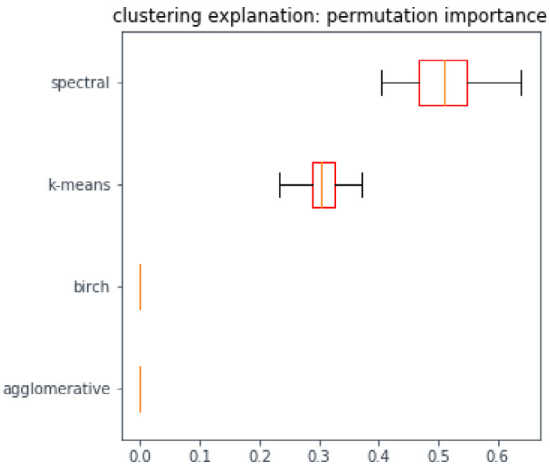

Fig. 6. Clustering importance: permutation importance.

Table 3
Clustering explanation: random forest.

Method	Importance
agglomerative	0.337020
k-means	0.298459
spectral	0.211150
birch	0.153371

Table 4
Clustering Explanation: LIME Global.

Method	Importance	Cluster
k-means	0.4772	0
k-means	−1.7429	1
spectral	−1.4031	0
spectral	0.6024	1

4.2. Algorithms explanation

Having used the stacking method in clustering therefore allows us to analyze and interpret each single method used as a feature, in order to see the global and local contribution of each method in order to create a robust and homogeneous grouping within and the most heterogeneous on the outside, between the instances present in the data. To do this are used the methods discussed, such as LIME, Random forest and Permutation importance.

From Fig. 6 it is possible to see that the two algorithms that have a greater weight are spectral clustering and *k*-means, and this is consistent with what was stated in the section dedicated to methodology, as by choosing a Euclidean kernel the spectral is reduced to *k*-means, which it has been seen to be one of the simplest and most transparent methods to use.

Regarding the random forest method for importance it can be seen in Table 3 that the most important methods are the agglomerative and the *k*-means, even here it is not surprising as the simplicity of the two methods indicates a clear transparency on how such. methods work.

Table 4 clearly shows the weight of each method on the global classification in clustering, also here with the LIME method it can be observed that *k*-means and spectral clustering are the most robust methods in the clustering of instances.

4.3. Features importance

For the part concerning the importance of each feature it can consider the application of some methods discussed above, one is certainly

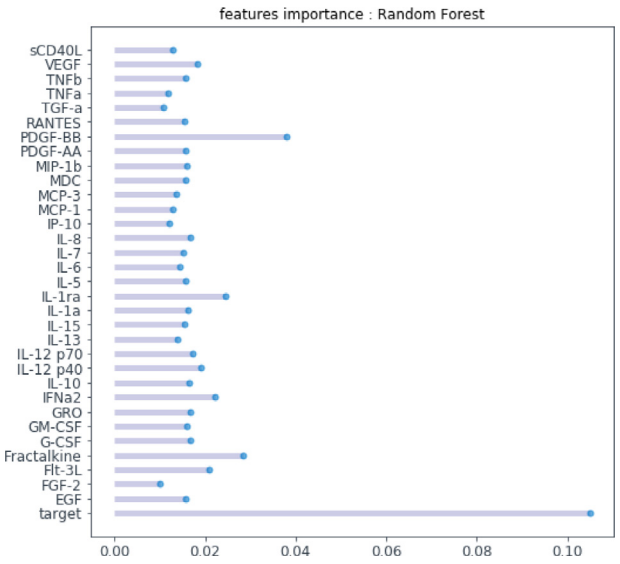

Fig. 7. Feature importance: random forest.

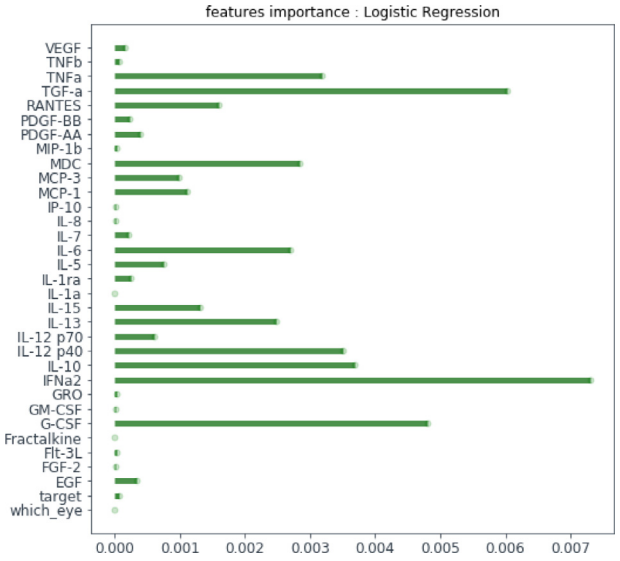

Fig. 8. Feature importance: logistic regression.

the one based on random forest and logistic regression, whose method is based on the standard deviation of the coefficients estimated by the model.

Fig. 7 shows the application of the random forest and Fig. 8 that of logistic regression, the most significant feature from the plot was omitted as its high value did not render the visualization of the other features well. Respectively in each method the value of the feature *which eye* (binary 0–1) obtained a score equal to 0.35 (random forest) and 5.80 (logistic regression), confirming the significance of the feature in both methods. As regards the impact of the other features, the variable named *target* (binary) which indicates the presence or absence of HIV in the subject, is confirmed as significant in the logistic regression and practically null in the random forest.

Fig. 9 shows the method known as permutation feature importance [35] which also shows that the variable which eye has a significant impact on the final classification, equal to 0.40, in accordance with random forest and logistic regression.

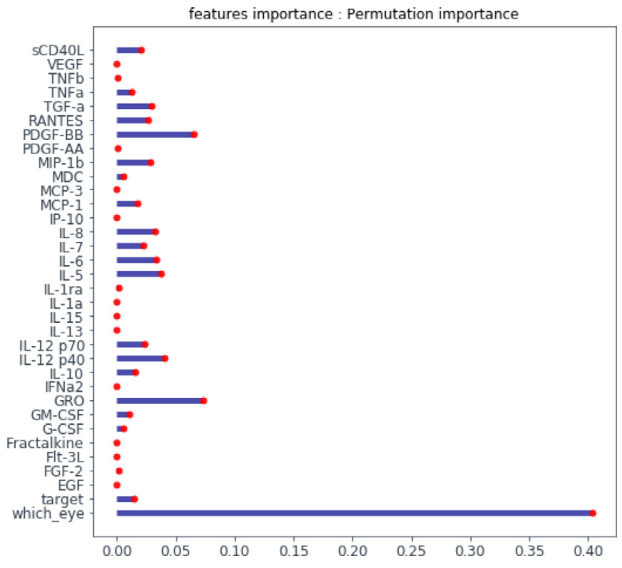

Fig. 9. Feature importance: permutation features importance.

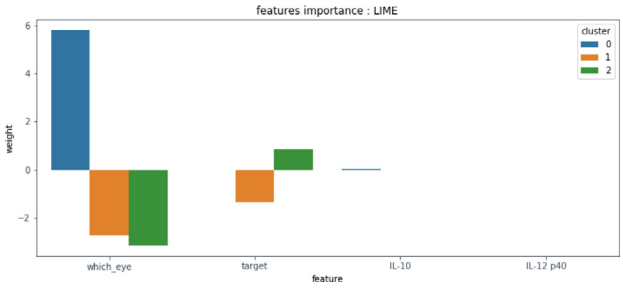

Fig. 10. Feature explainability: LIME.

4.4. Explainability

After providing a result in terms of the impact of the features on the classification, in this section are presented the results from the point of view of explainability and interpretation. Let us consider our logistic classifier and use a method known as LIME defined in the previous sections: Fig. 10 shows the explanation of the features at a global level, that is for the whole model, and it can be seen that also in this case, depending on the cluster, it can evaluate the single impact of the feature. In this case the significant impact is always the one that refers to the feature *which eye* which in cluster 0 increases the probability of belonging to that particular group by 6 times and in a minor (negative) but still significant form has an impact on the two remaining clusters, as proof and validation of the methods used in the previous subsection.

Using the LIME method, which provides a local interpretation, for a single instance, of how the features impact the final classification, Tables 5–7 show the values explained for three different patients. Each table shows the probability of belonging to a certain cluster for that patient, with the name of each feature and the corresponding value that affects that certain probability; in the case of patient CTH024 (Table 7) it is noted that the value of cytokine TGF-a has a strong contribution in identifying group 2 as probable to insert that patient, as well as in group 1 for patient CTH025 (Table 6) in which the values of the cytokines EGF and GRO have a strong relationship with belonging to that group.

5. Clinical explainability

This section provides an explanation of the results obtained above, from the clinical point of view and the possible implications, in order to provide the complete CDSS tool, assuming that the clinical operator works closely with the analytics expert and more generally of artificial intelligence, so that the results are robust from a methodological point of view, above all easily usable and interpretable in the clinical domain.

Table 5
LIME Explanation: patient: TH003.

y = 0 (p: 0.953)	Contribution	Feature
	+7.198	which eye
	+0.421	G-CSF
	+0.267	IL-7
	+0.236	EGF
	+0.217	GRO
	+0.157	IL-10
	+0.101	IL-15
	+0.100	IL-13
	+0.098	sCD40L
	−1.072	IP-10

Table 6
LIME Explanation: patient CTH025.

y = 1 (p: 0.027)	Contribution	Feature
	+0.924	GRO
	+0.713	G-CSF
	+0.608	IL-7
	+0.428	IL-10
	+0.302	sCD40L
	+0.260	EGF
	+0.241	IL-15
	+0.236	IL-8
	+0.179	IL-13
	−0.336	IP-10

Table 7
LIME Explanation: patient CTH024.

y = 2 (p: 0.887)	Contribution	Feature
	+5.535	TGF-a
	+1.326	PDGF-AA
	+1.091	IL-6
	+0.906	MIP-1b
	+0.869	HIV presence
	+0.829	IL-15
	+0.819	IL-5
	+0.734	IP-10
	+0.701	FGF-2
	+0.692	MDC

Table 8
LIME Explanation: patient TH003.

y = 1 (p: 0.95)	Negative	Positive	Feature	Value
0.46	0.00 < which eye		which eye	1.00
0.30	0.00 < HIV presence ≤ 1.00		HIV presence	1.00
0.16		TGF-a ≤ 503.15	TGF-a	105.72
0.10		FGF-2 ≤ 696.45	FGF-2	353.09
0.09	PDGF-BB ≤ 2309.25		PDGF-BB	1523.50
0.05		IP-10 ≤ 172976.24	IP-10	223207.87
0.07	GRO ≥ 20550.58		GRO	17005.63

5.1. Clustering explanations

In the results obtained by Agrawal et al. [19], it was highlighted that some cytokines are in close association with the DED pathology, therefore starting from their observations and their analysis results, in this phase of explainability of clustering, the results obtained on the basis of the cytokines GRO, EGF and IP-10, with respect to some features, such as which eye (1=right, 0=left) and the presence of HIV (0 = negative, 1 = positive) and obviously the label (obtained through the consent function) of membership in order to evaluate the assignment.

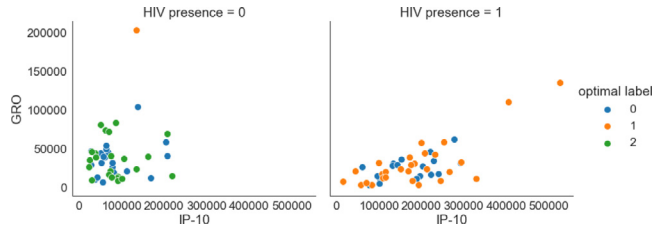

Fig. 11. Comparison IP-10 and GRO cytokines.

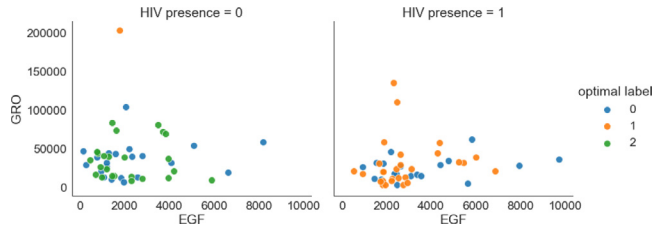

Fig. 12. Comparison EGF and GRO cytokines.

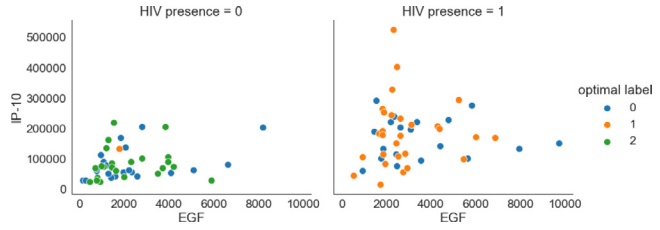

Fig. 13. Comparison IP-10 and EGF cytokines.

Figs. 11–13 show the clustering results with respect to the EGF, IP-10 and GRO features. We can observe that in Fig. 11, from the comparison between GRO and IP-10 the good ability of the method to insert in group 1 the seropositive patients (HIV = 1) emerges, equal to 60% of the instances and the remaining 40% are included in the cluster 0; in cluster 1 there are those patients who actually have very high IP-10 values. In cluster 2, on the other hand, it is possible to observe immunocompetent patients (equal to 50%) with higher GRO values and lower IP-10, confirming the results obtained by the authors [19]

By comparing the cytokines EGF and GRO, from Fig. 12 it note that here too it has an interesting result; in cluster 0, 4% of the units of this group have quite high values while those of the EGF are much lower in patients with HIV belonging to cluster 1. Fig. 13 confirms what has been said about the levels of IP-10 in relation to the EGF for patients in cluster 1 and 0.

5.2. Supervised explanations

Once it is obtained the labels are applied to a supervised binary classifier and based on the probabilities obtained for each patient, using the explainability methods 5–8. Obtaining the labels, it is applied to a supervised binary classifier and on the basis of the probabilities obtained for each patient, are used the explanability methods (Tables 5–8, obtaining some information of interest; for a given instance (patient: TH003, with HIV presence, 6) shows significant levels of EGF and GRO in order to increase the probability of belonging to cluster 1, that of probable patients with HIV and DED, while the cytokine IP-10 slightly decreases this impact. The eye-related feature has a significant contribution to the classification. For the patient CTH025 (immunocompetence, Table 6) with low probability of belonging to cluster 1 (in fact the hypotheses are that in cluster 1 there are HIV positive with DED), GRO, EGF and IP-10 also in this case are features of interest. In Table 7, the immunocompetent patient CHT024 with a high probability of belonging to group 2 (in which it is possible there are patients without HIV with DED, consistent since from the HIV presence features it is known that the patient is seronegative) so that the EGF and GRO are not the most influential while IP-10 is. In Table 8 for patient TH003, through the LIME method always with 95% probability of belonging to cluster 1, so that the values of the cytokine GRO equal to 17005.63, if higher than 20000, increase the risk by 7%, therefore consistent with the [19] results as below this threshold the risk decreases. The cytokine IP-10 confirms this hypothesis, since the present value for the instance is equal to 223207.87, a value less than 172976 would decrease the probability by 5%.

5.3. Extended discussion

In a clinical decision-making process, which makes use of techniques based on intelligent systems, the results obtained as those obtained in this work (Sections 4 and 5), without the use of explanability methods, could remain confined to a domain too technical related to the ability to understand the underlying mathematical method. Using the proposed methods, both for the decomposition of an ensemble method (in the case examined, clustering), and for the features importance and features explanation, the clinical operator is equipped with a complete tool able to collect the necessary data, analyze them, predict or classify a phenomenon, make it interpretable and transparent. The proposed framework could be configured in a broader framework of prescriptive analysis since starting from the inferences deduced within the data and from the predictions (related to the probability of developing the pathology) it is possible to make practical decisions about the problem treated. In the case of study treated, the results of Tables 5–8, for example, support the diagnosis of the clinical operator for a specific patient, for which it is possible to deduce both which are the cytokynes with abnormal values, and which suffering from DED (i.e. patient TH003). Also imagining a decision maker who obtaining an output from a model based on AI or ML, is induced to make a decision on how to approach the therapy for DED pharmacologically (for example) but without having evidence

1. on the characteristics (cytokines) that determine the increase or decrease in the risk probability
2. on the clinical picture of the pathology for each individual patient
3. on the impact of the disease in subjects who do not have similar characteristics

The proposed method would be able to identify the levels of cytokyne involved, associated with the probability estimated by the model (or models), the relationship between them and would allow the decision maker to have a detail on each component of the decision-making process. In the works mentioned in the first part of the work (Sections 1.1–1.2), these characteristics are not used, the works focus a lot on the accuracy of the models to predict the risk of disease onset, on the sensitivity of the models in recognizing this risk, but no one focuses on the interpretation of the model itself and above all, fundamentally, on what the model explains. Regarding the treatment and study of DED in HIV-infected patients, this work could constitute an extra step in the study of the disease and its treatment, placing itself in a field of literature in a middle way between machine learning and the clinical study of the phenomenon.

6. Conclusions

The use of a hybrid machine learning tool to support a clinical study such as the one treated has proven to be highly functional; combining unsupervised and supervised techniques, methods of features importance and explainable ML allowed us to build a robust tool that in a CDSS could be absolutely supportive. Logistic regression obtained an

accuracy of 91% on the train data and 86% on the test; the choice of logistic regression as meta-learner for the classification is motivated by the fact that the same authors [19] used logistic regression in their study, but obviously nothing prevents other methods from being used for this particular case of study (i.e. decision trees, neural network, . . .), but taking into account that using complex black-box methods such as neural networks, for example, always provides for the use of techniques for explaining the results as was done in this work by means of the LIME or Shapley method. The results obtained confirm the previous study by the [20] authors, [19], regarding the values of the cytokines GRO, EGF and IP-10 and their association with DED disease and seropositivity: this work adds a small contribution on how to use these [20] data, on how to interpret the results and another point of view on how to study the associated phenomenon.

6.1. Limitations and future work

By introducing this new methodology in a clinical decision-making process, decision makers will surely have an extra tool to deal with the diagnosis and treatment of this particular pathology that has been treated. However, there are several questions that at present can be investigated and further solutions to be pursued; for example if the proposed framework, based on clustering methods in which groups are chosen a priori, can be extended to the use of different methods, perhaps not necessarily based on Euclidean distances, or if textual data can be used for example, collected from medical records, or image data. In the hypotheses just made it is clear that the tools can be different and more complex, in the case study the data were numerical, extending to categorical, textual and image data, or in any case unstructured data, the resources to be put in place different. Instead of perhaps using a k -means method you will need to use a k -modes or other methods, such as autoencoder or self-organizing maps. Another consideration is related to data: how could the method behave if massive amounts of data were used? Therefore it should also be clear which and how many computational resources to put in place and if the explainability methods can be extended to unstructured data; fortunately, advances in explainable AI and clinical research continue and to date some of the questions posed have already been answered. The proposed method is applied to few data and computationally there were no difficulties and the mathematical methods used did not have any problems in use; in general, clinical problems have quite manageable datasets, since the aim of the work was to show both the potential of the proposed method and to make a contribution in the field of research for DED, although there are some limitations, the framework presented is in any case usable in the light of recent advances in this area.

Declaration of competing interest

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

References

[1] F.T. de Dombal, D.J. Leaper, J.R. Staniland, A.P. McCann, J.C. Horrocks, Computer-aided diagnosis of acute abdominal pain, 2(5804) (1972) 9–13, http://dx.doi.org/10.1136/bmj.2.5804.9.

[2] E. Shortliffe, Computer-based medical consultations: MYCIN, Artificial Intelligence 388 (1976) http://dx.doi.org/10.1097/00004669-197610000-00011.

[3] R.A. Miller, H.E. Pople, J.D. Myers, Internist-I, an experimental computer-based diagnostic consultant for general internal medicine, N. Engl. J. Med. 307 (8) (1982) 468–476, http://dx.doi.org/10.1056/NEJM198208193070803, PMID: 7048091.

[4] J. Müller, M. Stoehr, A. Oeser, J. Gaebel, M. Streit, A. Dietz, S. Oeltze-Jafra, A visual approach to explainable computerized clinical decision support, Comput. Graph. 91 (2020) 1–11, http://dx.doi.org/10.1016/j.cag.2020.06.004.

[5] H. Schafer, S. Hors-Fraile, R. Karumur, A. Calero Valdez, A. Said, H. Torkamaan, T. Ulmer, C. Trattner, Towards health (aware) recommender systems, 2017, http://dx.doi.org/10.1145/3079452.3079499.

[6] S. Tonekaboni, S. Joshi, M.D. McCradden, A. Goldenberg, What clinicians want: Contextualizing explainable machine learning for clinical end use, 2019, URL arXiv:1905.05134.

[7] Y. Xie, G. Gao, A. Chen, Outlining the design space of explainable intelligent systems for medical diagnosis, 2019.

[8] M. Naiseh, Explainability design patterns in clinical decision support systems, 2020.

[9] A. Bussone, S. Stumpf, D. O'Sullivan, The role of explanations on trust and reliance in clinical decision support systems, in: 2015 International Conference on Healthcare Informatics, 2015, pp. 160–169.

[10] M. Naiseh, N. Jiang, J. Ma, R. Ali, Explainable recommendations in intelligent systems: Delivery methods, modalities and risks, 2020.

[11] T. Liyuan, C. Zhang, L. Zeng, S. Zhu, N. Li, W. Li, H. Zhang, Y. Zhao, S. Zhan, H. Ji, Accuracy and effects of clinical decision support systems integrated with BMJ best practice-aided diagnosis: Interrupted time series study, JMIR Med. Inform. 8 (2020) e16912, http://dx.doi.org/10.2196/16912.

[12] N. Peiffer-Smadja, T. Rawson, R. Ahmad, A. Buchard, G. Pantelis, F.c.-X. Lescure, G. Birgand, A. Holmes, Machine learning for clinical decision support in infectious diseases: A narrative review of current applications, Clin. Microbiol. Infect. 26 (2019) http://dx.doi.org/10.1016/j.cmi.2019.09.009.

[13] N. Fitriyani, M. Syafrudin, G. Alfian, J. Rhee, HDPM: An effective heart disease prediction model for a clinical decision support system, IEEE Access 8 (2020) 133034–133050, http://dx.doi.org/10.1109/ACCESS.2020.3010511.

[14] J. Schwartz, A. Moy, S. Rossetti, N. Elhadad, K. Cato, Clinician involvement in research on machine learning–based predictive clinical decision support for the hospital setting: A scoping review, J. Am. Med. Inform. Assoc. 28 (2020) http://dx.doi.org/10.1093/jamia/ocaa296.

[15] S. Malik, N. Kanwal, M. Asghar, M. Ali, I. Karamat, M. Fleury, Data driven approach for eye disease classification with machine learning, Appl. Sci. 9 (2019) http://dx.doi.org/10.3390/app9142789.

[16] S. Nam, T.A. Peterson, A. Butte, K.Y. Seo, H.W. Han, Explanatory model of dry eye disease using health and nutrition examinations: Machine learning and network-based factor analysis from a national survey, JMIR Med. Inform. 8 (2020).

[17] B.N. Nguyen, A.W. Chung, E. Lopez, J. Silvers, H.E. Kent, S.J. Kent, L.E. Downie, Meibomian gland dropout is associated with immunodeficiency at hiv diagnosis: Implications for dry eye disease, Ocular Surf. 18 (2) (2020) 206–213, http://dx.doi.org/10.1016/j.jtos.2020.02.003.

[18] S.D. Mathebula, P.S. Makunyane, Ocular surface disorder among HIV and AIDS patients using antiretroviral drugs, Afr. Vis. Eye Health 88 (2019) 78(1), http://dx.doi.org/10.4102/aveh.v78i1.457.

[19] R. Agrawal, P.K. Balne, A. Veerappan, V.B. Au, B. Lee, E. Loo, A. Ghosh, L. Tong, S.C. Teoh, J. Connolly, P. Tan, A distinct cytokines profile in tear film of dry eye disease (DED) patients with HIV infection, Cytokine 88 (2016) 77—84, http://dx.doi.org/10.1016/j.cyto.2016.08.026.

[20] Dataset of tear film cytokine levels in dry eye disease (DED) patients with and without HIV infection, Data in Brief 10 (2017) 14–16, http://dx.doi.org/10.1016/j.dib.2016.11.027.

[21] G. Choi, J. Yun, J. Choi, D. Lee, J. Shim, H. Lee, Y.-H. Chung, Y. Lee, B. Park, N. Kim, K.M. Kim, Development of machine learning-based clinical decision support system for hepatocellular carcinoma, Sci. Rep. 10 (2020) 14855, http://dx.doi.org/10.1038/s41598-020-71796-z.

[22] E. Zihni, V.I. Madai, M. Livne, I. Galinovic, A.A. Khalil, J.B. Fiebach, D. Frey, Opening the black box of artificial intelligence for clinical decision support: A study predicting stroke outcome, PLoS One 15 (2020) 1–15, http://dx.doi.org/10.1371/journal.pone.0231166.

[23] E. Georga, V. Protopappas, E. Arvaniti, D. Fotiadis, The Diabino System: Temporal Pattern Mining from Diabetes Healthcare and Daily Self-monitoring Data: ICBHI 2015, Haikou, China, 8-10 October 2015, 2019, pp. 61–65, http://dx.doi.org/10.1007/978-981-10-4505-9_10.

[24] E. Kumar, P. Jayadev, Deep Learning for Clinical Decision Support Systems: A Review from the Panorama of Smart Healthcare, 2020, pp. 79–99, http://dx.doi.org/10.1007/978-3-030-33966-1_5.

[25] F. Curia, Explainable Clinical Decision Support System: Opening Black-Box Meta-Learner Algorithm Expert's Based (Ph.D thesis), Catalogo Iris, Sapienza University of Rome, 2021, URL http://hdl.handle.net/11573/1538472.

[26] A. Barredo Arrieta, N.D. az Rodrí guez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Inf. Fusion 58 (2020) 82–115, http://dx.doi.org/10.1016/j.inffus.2019.12.012.

[27] M.T. Ribeiro, S. Singh, C. Guestrin, Why should I trust you?: Explaining the predictions of any classifier, 2016, URL arXiv:1602.04938.

[28] Z.C. Lipton, The mythos of model interpretability, 2017, URL arXiv:1606.03490.

[29] P. Smyth, D. Wolpert, Linearly combining density estimators via stacking, Mach. Learn. 36 (1999) 59–83, http://dx.doi.org/10.1023/A:1007511322260.

[30] J.H. Friedman, Greedy function approximation: A gradient boostingmachine, Ann. Statist. 29 (5) (2001) 1189–1232, http://dx.doi.org/10.1214/aos/1013203451.

[31] A. Goldstein, A. Kapelner, J. Bleich, E. Pitkin, Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation, 2014, URL arXiv:1309.6392.
[32] J.H. Friedman, B.E. Popescu, Predictive learning via rule ensembles, Ann. Appl. Stat. 2 (3) (2008) 916–954, http://dx.doi.org/10.1214/07-AOAS148.
[33] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, Vol. 30, Curran Associates, Inc., 2017.
[34] P.W. Koh, P. Liang, Understanding black-box predictions via influence functions, 2020, URL arXiv:1703.04730.
[35] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32, http://dx.doi.org/10.1023/A:1010933404324.
[36] T.D. Gedeon, Data mining of inputs: Analysing magnitude and functional measures, Int. J. Neural Syst. 8 (2) (1997) 209–218, http://dx.doi.org/10.1142/S0129065797000227.
[37] A. "Zien, N. Krämer, S. Sonnenburg, G. Ratsch, The feature importance ranking measure, in: Machine Learning and Knowledge Discovery in Databases, Springer Berlin Heidelberg, "2009, pp. 694–709.
[38] A. Altmann, L. Tolosi, O. Sander, T. Lengauer, Permutation importance: a corrected feature importance measure, Bioinformatics 26 10 (2010) 1340–1347.
[39] A. Rosenberg, J. Hirschberg, V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure, 2007, pp. 410–420.
[40] J.C. Dunn, A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, J. Cybern. 3 (3) (1973) 32–57, http://dx.doi.org/10.1080/01969727308546046.
[41] P.J. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. 20 (1987) 53–65, http://dx.doi.org/10.1016/0377-0427(87)90125-7.
[42] T. Fawcett, An introduction to ROC analysis, in: ROC Analysis in Pattern Recognition, Pattern Recognit. Lett. 27 (8) (2006) 861–874, http://dx.doi.org/10.1016/j.patrec.2005.10.010.