Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

ICU admission and mortality classifiers for COVID-19 patients based on subgroups of dynamically associated profiles across multiple timepoints

Vasileios C. Pezoulas [a], Konstantina D. Kourou [a], Eugenia Mylona [a], Costas Papaloukas [a,b], Angelos Liontos [c], Dimitrios Biros [c], Orestis I. Milionis [c], Chris Kyriakopoulos [d], Kostantinos Kostikas [d], Haralampos Milionis [c], Dimitrios I. Fotiadis [a,e,*]

[a] *Unit of Medical Technology and Intelligent Information Systems, Dept. of Materials Science and Engineering, University of Ioannina, Ioannina, GR45110, Greece*
[b] *Dept. of Biological Applications and Technology, University of Ioannina, Ioannina, GR45100, Greece*
[c] *Dept. of Internal Medicine, School of Medicine, University of Ioannina, Ioannina, GR45110, Greece*
[d] *Respiratory Medicine Dept., School of Medicine, University of Ioannina, Ioannina, GR45110, Greece*
[e] *Institute of Biomedical Research, FORTH, Ioannina, GR45110, Greece*

ARTICLE INFO

Keywords:
COVID-19
Dynamic Bayesian Networks (DBNs)
Self-Organizing Maps (SOMs)
ICU admission
Mortality

ABSTRACT

The coronavirus disease 2019 (COVID-19) which is caused by severe acute respiratory syndrome coronavirus type 2 (SARS-CoV-2) is consistently causing profound wounds in the global healthcare system due to its increased transmissibility. Currently, there is an urgent unmet need to identify the underlying dynamic associations among COVID-19 patients and distinguish patient subgroups with common clinical profiles towards the development of robust classifiers for ICU admission and mortality. To address this need, we propose a four step pipeline which: (i) enhances the quality of multiple timeseries clinical data through an automated data curation workflow, (ii) deploys Dynamic Bayesian Networks (DBNs) for the detection of features with increased connectivity based on dynamic association analysis across multiple points, (iii) utilizes Self Organizing Maps (SOMs) and trajectory analysis for the early identification of COVID-19 patients with common clinical profiles, and (iv) trains robust multiple additive regression trees (MART) for ICU admission and mortality classification based on the extracted homogeneous clusters, to identify risk factors and biomarkers for disease progression. The contribution of the extracted clusters and the dynamically associated clinical data improved the classification performance for ICU admission to sensitivity 0.83 and specificity 0.83, and for mortality to sensitivity 0.74 and specificity 0.76. Additional information was included to enhance the performance of the classifiers yielding an increase by 4% in sensitivity and specificity for mortality. According to the risk factor analysis, the number of lymphocytes, SatO2, PO2/FiO2, and O2 supply type were highlighted as risk factors for ICU admission and the percentage of neutrophils and lymphocytes, PO2/FiO2, LDH, and ALP for mortality, among others. To our knowledge, this is the first study that combines dynamic modeling with clustering analysis to identify homogeneous groups of COVID-19 patients towards the development of robust classifiers for ICU admission and mortality.

1. Introduction

Among the infected individuals with SARS-CoV-2 [1], it is estimated that 1/3 of them never develop symptoms [2,3] and those who will develop symptoms may have a mild to moderate self-limiting disease [3]. In contrary, the severity of symptomatic infection ranges from mild to critical, and most individuals will develop a non-severe illness [4]. The progression of the disease and the risk of severe illness varies by age, underlying comorbidities, and risk factors for disease progression, such as, cardiovascular diseases, diabetes mellitus, chronic obstructive pulmonary disease, cancer (e.g., hematologic malignancies, lung cancer), chronic kidney disease, solid organ or hematopoietic stem cell transplantation, obesity, and smoking [5]. According to the official report of the Centers for Disease Control and Prevention (CDC) in the US, among 1.3 million confirmed COVID-19 cases, 14% of patients were hospitalized, 2% were admitted in the intensive care unit (ICU), and 5% died

[*] Corresponding author. Unit of Medical Technology and Intelligent Information Systems, Dept. of Materials Science and Engineering, University of Ioannina, GR45110, Ioannina, Greece.
E-mail address: fotiadis@uoi.gr (D.I. Fotiadis).

https://doi.org/10.1016/j.compbiomed.2021.105176
Received 11 November 2021; Received in revised form 22 December 2021; Accepted 23 December 2021
Available online 27 December 2021

[6]. In addition, the risk of critical or fatal disease is high among hospitalized COVID-19 patients [7,8].

The increased need for intensive care units and ventilators due to the unprecedented number of confirmed COVID-19 cases has surpassed the capacity of international healthcare systems. As a result, the World Health Organization (WHO) highlighted the importance of artificial intelligence (AI) as a prominent solution to manage the crisis caused by the virus [9]. AI is a constructive, non-medical intervention approach with a strong potential to overcome the current global health crisis, build next-generation epidemic preparedness, and move towards a resilient recovery [9]. Moreover, AI can shed light into the clinical unmet needs in COVID-19, including the development of robust models for: (i) the prediction of ICU admission, mortality, and the need for mechanical ventilation, (ii) the extraction of prominent risk factors for ICU submission and mortality, (iii) the early suggestion of targeted interventions/therapeutic treatments, and (iv) the definition of better disease severity indices. Although AI is a promising tool to unveil the underlying mechanisms of COVID-19, the risk of bias and discrimination in its design and deployment must be taken into consideration.

According to the literature, several studies have deployed AI to address the clinical unmet needs in COVID-19. Bagging methods, such as, the Random Forest algorithm, were used for risk stratification based on time-series data across 1987 unique patients diagnosed with COVID-19 and admitted to non-ICU units to optimize the flow of operations within the hospitals [10]. Bagging methods have also been applied on clinical data from 362 patients with confirmed COVID-19, highlighting age, hypertension, gender, absolute neutrophil count, IL-6, and LDH as risk factors for disease severity [11]. Ensemble-based algorithms, such as, the gradient boosting trees, have been widely used to predict 5-day ICU admission and 28-day mortality across 3597 COVID-19 patients, stressing the importance of CRP, LDH, and O2 saturation for ICU admission and neutrophil and lymphocyte percentages for mortality [12]. Ensemble learning has been deployed to identify an optimal combination of factors that predicts ICU admissions across 733 patients diagnosed with COVID-19 [13], as well as, across 1270 COVID-19 patients [14], highlighting the age, CRP, and LDH as prominent features for mortality. Furthermore, multipurpose machine learning algorithms (e.g., artificial neural networks and ensemble classifiers) have been proposed to estimate the risk of ICU admission or mortality among 3623 hospitalized patients with COVID-19, yielding a good discrimination performance [15], as well as, across 3280 patients to predict the risk of developing critical conditions in COVID-19 with high predictive performance [16].

Nonetheless, none of these studies have thoroughly investigated the dynamic associations among clinical, laboratory and biological data across multiple time intervals nor they have shed light into the interpretability and explainability of the risk predictors for ICU admission and/or mortality of hospitalized COVID-19 patients. Furthermore, none of the existing studies focus on the development of data curation workflows to improve the quality of the available clinical, laboratory and biological data across multiple time-points. This is a major concern, since data with insufficient quality stemming from the hospital crisis may hamper the effective management of COVID-19. Moreover, the application of clustering and trajectory analysis to extract homogeneous groups of COVID-19 patients with common clinical profiles are two promising approaches that may further enhance the predictive value of the AI models for ICU admission and mortality. As a matter of fact, the lack of ICU admission and mortality classifiers which take into consideration the underlying dynamic associations to identify homogeneous clusters of COVID-19 patients, remains an unmet need.

To address these needs, we propose a pipeline which: (i) utilizes dynamic modeling approaches to extract highly associated features across multiple time-points, (ii) uses these features to extract clusters of COVID-19 patients with common clinical profiles, (iii) combines the results from the clustering analysis and the dynamic modeling process to develop robust classifiers for ICU admission and mortality, (iv) enhances the performance of the classifiers using baseline clinical data, therapies and demographics, and (v) identifies prominent risk factors for ICU admission and mortality. More specifically, Dynamic Bayesian Networks (DBNs) are used to capture the features having the highest degree and connectivity across multiple time-points within a directed acyclic graph. Self-Organizing Maps (SOMs) are trained on the highly associated features used to extract homogeneous clusters of patients with common clinical course. The extracted clusters are combined with the high-quality time-series clinical data to develop robust classifiers for ICU admission and mortality. In this way, the features having the highest degree of connectivity in the extracted DBN are used to extract homogeneous clusters of COVID-19 patients with common clinical profiles based on the SOMs (and the trajectories) to enhance the robustness of the classifiers for ICU admission and mortality.

Three case studies were conducted to evaluate the performance improvement in classifying the patient subgroups derived from the SOMs. Our results highlight the contribution of the extracted patient subgroups in the improvement of the classification performance for ICU admission up to sensitivity 0.83 and specificity 0.83, and for mortality up to sensitivity 0.74 and specificity 0.76. Additional baseline data were included in the input space to improve the performance of the classifiers, yielding an increase of 4% in sensitivity and specificity for ICU admission and 3% in sensitivity and 2% in specificity. The risk factor analysis highlighted the number of lymphocytes, SatO2, PO2/FiO2, and O2 supply type as risk factors for ICU admission and the percentage of neutrophils and lymphocytes, PO2/FiO2, LDH, and ALP for mortality.

The paper is structured as follows. Section 2 offers a comprehensive view on the methods which were utilized in the current study, including: (i) time-series data curation, (ii) dynamic association analysis based on DBNs, (iii) clustering analysis based on SOMs and Latent Growth Mixture Modelling (LGMM), (iv) classifiers for ICU admission and mortality with class imbalance handling, and (v) risk factor analysis. The results of the overall analysis are presented in Section 3, including the inferred DBN, the homogeneous patient clusters and the identified risk factors. The outcomes are discussed in Section 4 and future work in Section 5.

2. Materials and methods

2.1. Data origin and problem definition

Anonymized baseline and follow up clinical data (Supplementary Table 1) were acquired from the Dept. of Internal Medicine at the University Hospital of Ioannina. In total, 422 hospitalized COVID-19 patients were included in the analysis with an average age of 64.28 (±16.72) years. The time-series data consisted of 51 clinical features across 7 timepoints: 1, 3, 5, 7, 9, 11, and 15 days after hospitalization. Out of 422 patients, 25 patients (5.92%) were admitted in the ICU and 49 patients died (11.61%). Out of the 49 patients who died, 18 were admitted in the ICU. The classification tasks are formulated as follows: (i) in the first case, the target group consists of the patients who were admitted in the ICU (25 patients), and (ii) in the second case, the target group consists of the patients who died (49 patients). In each case, the remaining patients are assigned to the control group.

2.2. Workflow overview

According to Fig. 1, the overall workflow consists of four steps, including: (i) time-series data curation in order to enhance the quality of the available time-series data by automatically removing data inconsistencies and applying data-driven imputation methods by taking into consideration the neighboring clinical profiles for each missing record, (ii) dynamic association analysis for the identification of features with increased connectivity through the application of DBNs, (iii) SOMs and trajectory analysis for the extraction of homogeneous clusters of patients with common clinical profiles, and (iv) the application of

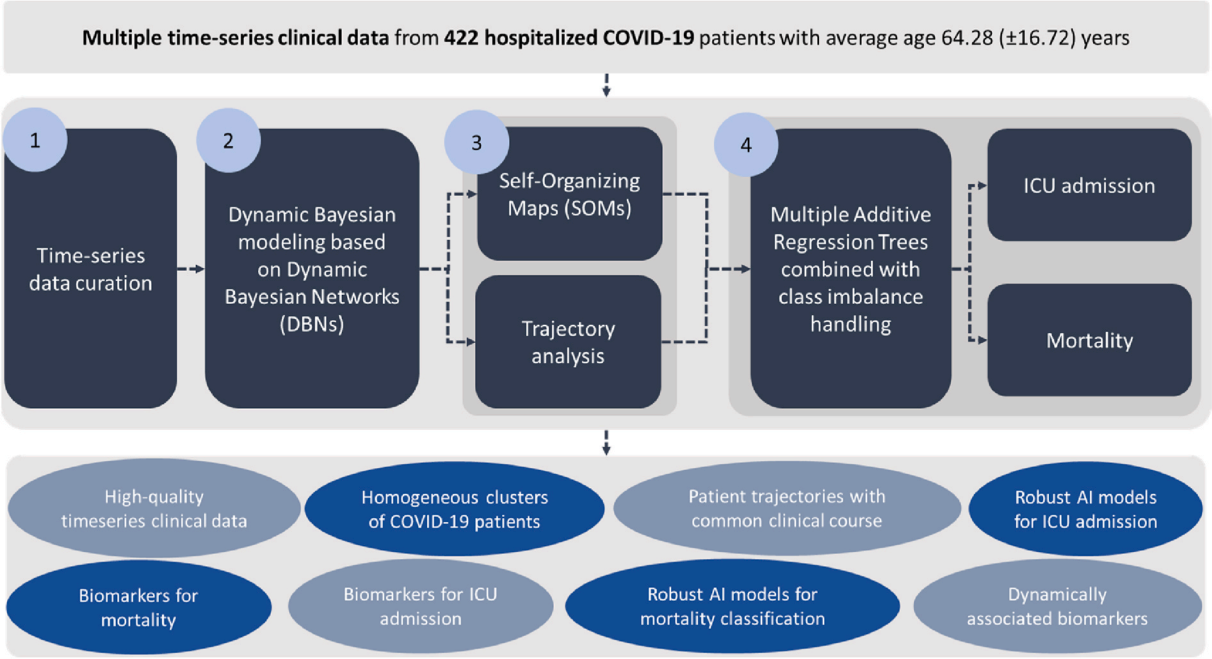
Multiple time-series clinical data from 422 hospitalized COVID-19 patients with average age 64.28 (±16.72) years
1
Time-series data curation
2
Dynamic Bayesian modeling based on Dynamic Bayesian Networks (DBNs)
3
Self-Organizing Maps (SOMs)
Trajectory analysis
4
Multiple Additive Regression Trees combined with class imbalance handling
ICU admission
Mortality
High-quality timeseries clinical data
Homogeneous clusters of COVID-19 patients
Patient trajectories with common clinical course
Robust AI models for ICU admission
Biomarkers for mortality
Biomarkers for ICU admission
Robust AI models for mortality classification
Dynamically associated biomarkers

Fig. 1. Illustration of the workflow analysis.

Multiple Additive Regression Trees (MART) combined with class imbalance handling for ICU and mortality classification across three different case studies, which involve: (a) the 51 clinical and laboratory features across the first 4 timepoints with and without the clustering labels from SOMs (case study 1), (b) the most important features from the DBNs across the first 4 timepoints with and without the clustering labels from SOMs (case study 2), and (c) only the clustering labels from the SOMs (case study 3). The three case studies were employed to investigate whether clustering analysis can contribute to the performance of the classifiers for ICU admission and mortality. The outcomes of the workflow include, besides the homogeneous SOMs clusters and the trajectories of COVID-19 patients, high quality time-series clinical data, dynamically associated biomarkers for ICU admission and mortality, and robust AI models for ICU admission and mortality classification.

2.2.1. *Time-series data curation*

The data curation pipeline presented in Pezoulas et al. [17], was extended to support the analysis of time-series clinical data. The latter were categorized according to their quality into three states, namely the “good”, “fair” (<30% missing values) and “bad” (>30% missing values), where the “bad” features were discarded from further analysis. Multivariate methods were used to isolate outliers. Data imputation was applied on the “fair” data based on the k-nearest neighbors (kNN) approach, where information from the clinical profiles of the neighboring patients was used to fill in the missing values.

2.2.2. *Dynamic Bayesian Networks (DBNs) for feature selection*

Bayesian Networks (BNs) refer to the general class of graphical models in which nodes and the edges between them denote the assumptions on their conditional dependence [18]. Although probability and conditional independencies characterize BNs, the concept of causal influence can also be defined [19]. It is possible to identify causal reasoning (from known causes to unknown effects) and/or diagnostic reasoning (from known effects to unknown causes) in a BN. In the present study, a DBN model has been designed and developed to represent the conditional dependencies over time (four discrete time-points) for certain variables (i.e., clinical, therapies, laboratory-related). DBNs, as an extension of BNs, enable the: (i) modeling of stochastic phenomena, (ii) incorporation of prior knowledge, and (iii) handling of hidden variables [20]. They have been used for discovering how a random variable X evolves over time during a stochastic process [20,21].

DBNs are defined by a graphical structure and a set of parameters. DBN theory is generally based on two assumptions [22]. First, the process is Markovian in the set of variables X, i.e., $P(X[t+1]|X[0],\ldots,X[t]) = P(X[t+1]|X[t])$. Second, it is assumed that the process is stationary, i.e. the transition probability $P(X[t+1]|X[t])$ is independent of t. To represent beliefs about the possible trajectories of the process, we need a probability distribution over random variables for all t. A DBN that represents the joint distribution over all possible trajectories consists of two parts [22]: (i) a prior network B_0 that specifies a distribution over initial states $X[0]$, and (ii) a transition network, $B_{\rightarrow}$, over the variables $X[0] \cup X[1]$ that is taken to specify the transition probability $P(X[t+1]|X[t])$ for all t.

Given a DBN model, the joint distribution over $X[0],\ldots,X[T]$ is:

$$P_B(x[0],..,x[T]) = P_{B_0}(x[0]) \Pi_{t=o}^{T-1} P_{B_{\rightarrow}}(x[t+1]|x[t]). \qquad (1)$$

The implementation of the Dynamic Bayesian Networks (DBNs) was conducted using the “bnstruct” package [23] in R4.0.3 for learning the structure and the parameters of the network, where four time slices were considered to calculate the joint distribution probabilities over time for the continuous features. Learning the structure of a DBN corresponds to the specification of the intra-slice and the inter-slice topologies. In addition, the conditional probability distributions (CPDs) at each node were computed. The parameters specified for structure learning were: (i) the time-series clinical data, (ii) the state of variables (discrete or continuous), (iii) the names of the variables, (iv) the number of levels they must be quantized into (in our case equals to 4 according to their variance), and (v) the number of time-points. The “ggplot2” package [24] was used to depict the structure of the DBN (Fig. 3).

2.2.3. *Cluster analysis*

2.2.3.1. Self-Organizing Maps (SOMs). SOMs adopt a competitive learning strategy according to which low dimensional projections of high-dimensional input data are generated by a sequential training process [25,26]. The latter utilizes a SOM grid (e.g., a rectangular grid) on top of which the weight vectors of a pre-defined number of neurons is adjusted by computing the Euclidean distance between the input samples and the neurons in the grid. Then, the neurons are re-adjusted in the grid and the neuron with the smallest Euclidean distance is extracted as the best matching unit (BMU) according to the following weight update function:

$$w_j(n+1) = w_j(n) + Q(j, \mathrm{k}, n)\beta(n)\big(X(i) - w_j(n)\big), \tag{3}$$

where w_x is the weight vector of neuron j, n is the iteration stage, i is the index of the input vector, k is the index of the BMU, $X(i)$ is the i-th input vector, $\beta(n)$ is a learning coefficient, and $Q(j,k,n)$ is a neighborhood function which calculates the distance between neurons j and k, at step n. The SOMs clusters were associated with ICU admission and mortality using the Fisher's exact test [27]. The implementation of the SOMs took place in R4.0.3 using the "SOMbrero" package [28]. A 7x7 square grid topology was utilized for the training process, where the Euclidean distance was used to define the topology of the grid. An aggregation process was finally applied based on hierarchical clustering to further combine the individual clusters yielding a final set of 4 superclusters for each feature.

2.2.3.2. Trajectories. For each one of the features found to have increased connectivity in the DBNs, Latent Growth Mixture Modelling (LGMM) was performed to identify underlying clusters of individual trajectories in the studied population. LGMM is a data-driven processes that combines latent growth curve and mixture models, where the latent classes, or clusters, in the population are estimated by probabilistically grouping individuals with similar starting points (intercepts) and patterns of change (slopes). The advantage of LGMM is that it allows to estimate within-class variability for each individual trajectory which describes how closely individuals within a class resemble the mean trajectory [29]. As described in Ref. [30], a LGMM model can be written as:

$$Y[t]_n = \sum_{c=1}^{C}\big(\pi_{nc}\big(g_{0nc}\cdot A_{0c}[t] + g_{1nc}\cdot A_{1c}[t] + e[t]_{nc}\big)\big), \tag{2}$$

given that:

$$0 \le \pi_{nc} \le 1 \text{ and } \sum_{c=1}^{C}(\pi_{nc}) = 1,$$

where the observed longitudinal data for the individual n on the left side of the equation (individuals' scores on variable Y repeatedly measured at times $t = 0$ to T) are represented using two latent variables, g_{0nc} and g_{1nc}, two corresponding basis vectors (i.e., sets of factor loadings), $A_{0c}[t]$ and $A_{1c}[t]$ describing the patterns or shapes of changes, and a time-specific residual (i.e., error), $e[t]_{nc}$. Possible differences among groups (or classes) are indicated by the c subscripts while π_{nc} is the probability of individual n to belong to class c.

For each feature, a series of models were fitted for 2 to 6 cluster solutions. To select the best optimal clustering solution, we derived a combination of fit statistics, including: (i) the most commonly used log-likelihood fit index, the Bayesian information criterion (BIC), where lower values indicate a better model fit [31], and (ii) two classification-based fit statistics: the scaled entropy, a measure of classification quality that ranges from 0 to 1 with higher values indicating more distinct classes [32], and the average posterior probability of assignment (APPA) (class-specific fit index), which is calculated as the average posterior probability of belonging to class k over all the individuals assigned to class k [33]. APPA is also bounded by 0 and 1 and ideally should exceed a minimum threshold value of 0.7 [34]. Apart from fit statistics, other factors were also considered. Clustering solutions that resulted to a class size comprising less than 5% of the sample were excluded to prevent overfitting. The classes interpretation and clinical meaningfulness was assessed through the plotting of group trajectories [30].

To approximate the true distribution function, both linear and nonlinear (i.e., beta cumulative distribution function and quadratic I-splines) link functions were considered, and their acceptability was determined using the discrete log-likelihood and the derived Akaike information criterion (AIC).

The "lcmm" package from R4.0.3 was used to fit the LGMM [35] and the "ggplot2" [24] to plot the trajectories (Fig. 5). In the "lcmm" function, we specified class-specific fixed effects of time on the trajectories, as well as, a random intercept and a random effect on time. The link function for fitting the model was either "beta" or "splines". As a final step, we investigated how the clusters for each feature were associated with ICU admission and mortality, using the Fisher's exact test.

2.2.4. Aggregation of DBNs with SOMs and trajectories

The M-best features across the first four timepoints from the DBNs are grouped into a set of best features, say $X_{DBN} = \{X_1, X_2, \dots X_M\}$. This set of features is then utilized in the SOMs to extract homogeneous clusters of COVID-19 patients with common clinical profiles. An individual SOM is trained on the time-series data from each feature in X_{DBN}, say X_i, $i \in [1, M]$ yielding a new feature with K-clustering labels, say X'_i, $i \in [1, M]$, where $X'_i \in [1, K]$. Each feature X'_i, $i \in [1, M]$ is utilized into a $X_{SOM} = \{X'_1, X'_2, \dots X'_K\}$. Then, the features from the DBNs are aggregated with the new features from the SOMs to investigate whether this aggregation enhances or not the performance of the ICU and mortality classifiers. The same procedure was applied for the clusters from the trajectory analysis.

2.2.5. Classification models for ICU admission and mortality

2.2.5.1. Multiple additive regression trees (MART). Multiple additive regression trees [36] combine a set of weak regression trees (learners) into a robust classifier through a series of sequential boosting stages, where on each stage the algorithm minimizes the gradient of a loss function to reduce the classification error. Here, we use the Gradient Boosting Trees (GBTs) classifier as a widely used type of MART. At step n, the GBTs algorithm seeks for a weak learner, say $m_i(n)$, which minimizes the following cost function:

$$M_n(x) = M_{n-1}(x) + argmin_f\left(\sum_{z=1}^{Z} L\big(y_z, M_{n-1}(x_z) + m_n(x_z)\big)\right), \tag{4}$$

where $L(.)$ is the error loss function, Z is the number of samples and y_z is the predicted value at step z. In the GBTs configuration schema, the booster was set to the gradient boosting trees followed by a random sampling of the training instances prior to the construction of trees to prevent overfitting. The Gradient Boosting classifier from the "scikit-learn" package was used for the development of the ICU admission and mortality classifiers based on regression tree learners, with learning rate 0.1, negative binomial log-likelihood loss function for binary classification tasks, 100 boosting stages, and a subsample ratio 0.9 (the fraction of samples to be used for fitting the weak tree learners).

2.2.5.2. Class imbalance handling. The number of patients who were admitted in the ICU (target group 1) was 25 (5.92%) whereas the number of patients who did not survive (target group 2) was 48 (11.37%). To deal with the increased class imbalance present in both target groups, random downsampling with replacement was applied to match each target group with the corresponding control group. The process was repeated 100 times to cover the whole population [37]. In each iteration, the downsampled controls were matched with the corresponding target populations according to age. A 10-fold stratified cross validation procedure was applied in each downsampling iteration to estimate the accuracy, sensitivity, specificity, and area under the ROC curve (AUC) of the classifiers for ICU admission and mortality. Finally, the performance evaluation results were averaged across the folds and the downsampling iterations.

2.2.6. *Risk factor analysis*

The F-score method was used to quantify the importance of each feature during the decision-making process, where the F-score of the *i*-th feature, say F_i, is defined as in Ref. [36]:

$$F_i = \frac{\left(\widehat{x}'_i - \widehat{x}_i\right)^2 + \left(\widehat{x}''_i - \widehat{x}_i\right)^2}{\frac{1}{n'-1}\sum_{j=1}^{n'}\left(x'_{j,i} - \widehat{x}'_i\right)^2 + \frac{1}{n''-1}\sum_{j=1}^{n''}\left(x''_{j,i} - \widehat{x}''_i\right)^2}, \tag{5}$$

where $\widehat{x}_i$, $\widehat{x}'_i$, $\widehat{x}''_i$ are the average values of the *i*-th feature in the whole, in the positive (i.e., positive target outcome), and in the negative (i.e., negative target outcome) datasets, respectively, $x'_{j,i}$ is the *j*-th positive instance of the *i*-th feature, $x''_{j,i}$ is the *j*-th negative instance of the *i*-th feature, n' is the number of positive instances, and n'' is the number of negative instances.

3. Results

3.1. *Time-series data quality*

The number of features with either good or fair quality status was 70 in timepoint 1, 66 in timepoints 1–2, 55 in timepoints 1–3, 51 in timepoints 1–4, 48 in timepoints 1–5, 28 in timepoints 1–6, and 20 in timepoints 1–7, where the time-points refer to hospitalization days. Consequently, only the 51 features having either fair or good quality status in timepoints 1–4 were considered as eligible for the analysis since the inclusion of information from additional timepoints would result in information loss due to the bad quality status. An overall description of the quality of the eligible features across the seven time-points is presented in Supplementary Table 2 whereas the quality status for each one of the 51 eligible features (32 continuous, 19 discrete) is summarized in Supplementary Table 3. According to Supplementary Table 2, the number of discrete features was 19 whereas the number of continuous features was 32. In both cases, the quality of the features is considered as adequate for the analysis until the 4th day of hospitalization. Out of the 32 continuous features (Fig. 2(A)), 9.37% was good, 65.18% was fair and 25.45% was bad whereas out of 19 discrete features (Fig. 2(B)), 15.78% were good, 57.9% were fair and 26.32% were bad, on average, across the available time-points. Data imputation based on the kNN approach was only applied for the features with fair quality. The abbreviations for the input features are presented in Supplementary Table 1.

3.2. *DBNs analysis results*

The 32 continuous features from timepoints 1–4 were utilized in the DBN analysis. Fig. 3 illustrates the chord diagram related to the adjacency matrix obtained after training the DBN model. This diagram permits the study of flows between the set of nodes within the network. The nodes are displayed all around a circle and connected with arcs (links). Chord diagrams are built from the adjacency matrix within a DBN and display the circular visualization of relations between nodes by links. In an adjacency matrix, value in i^{th} row and j^{th} column represents the relation between the object in row i^{th} to the object in the j^{th} column where the absolute value measures the strength of the relation. In an adjacency list, relations are represented as a three-column data frame in which relations come from the first column to the second column, and the third column represents the strength of the relation.

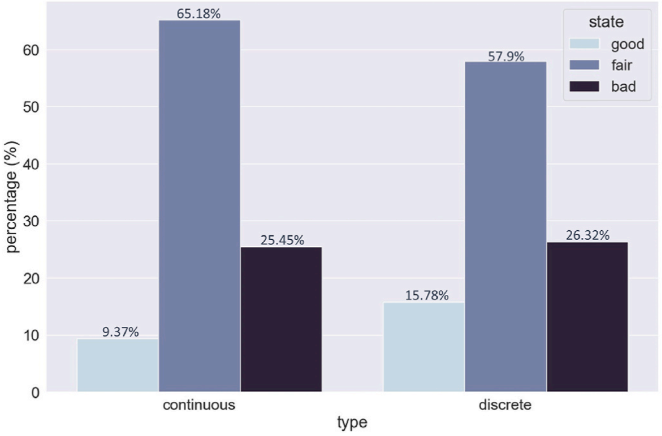

Fig. 2. Illustration of the quality status across the time-points for the continuous and the discrete features.

More specifically, Fig. 3 illustrates the relationships (links) found by the DBN analysis based on the time series data. An adjacency matrix was used to exploit the connections between the nodes. Each link between two nodes represents the calculated probability which reveals the inter-slice connection in a DBN model. In this context, we can extract information regarding the connection among the important clinical variables (i.e., the probabilistic inference among the set of variables, modeled using a directed acyclic graph), over time, regarding ICU admission and mortality.

We can thereby conjecture about the nodes that have the higher number of connections within the network model. Based on this knowledge, we observe that the absolute number of neutrophils has the higher degree of inter-relationships in the proposed model along with the cardiac frequency at day 1 when ICU admission and mortality classification of COVID-19 patients is considered. Hence, we anticipate this factor to be of high significance for disease prognosis as regards to ICU admission and patient risk stratification. Fig. 4 provides a more thorough explanation of the proposed DBN model (i.e., structure learning) about the centrality measures of each node regarding: (i) the in-degree, (ii) the out-degree, and (iii) the betweenness centrality of each entity. These measures were extracted based on the inferred DBN model and were grouped for each node.

The node degree corresponds to the number of connections for each node, while the betweenness measure refers to the node's influence over information flow. In the y-axis, the z-scores (i.e., standardized coefficients) are given allowing to understand the inter-connectedness of each node. Neutrophil number (“Neut_abs_no”) and cardiac frequency (“cardiac_freq”) at day 1 are two of the most interconnected variables based on their in- and out-degrees, allowing for better understanding of their links to illness progression and ICU admission at different timepoints.

3.3. *Cluster analysis results*

3.3.1. *SOMs super-clusters*

A 7x7 grid was utilized for the neuron training process (Section 2.2.3.1). The latter was applied on the 32 continuous features from Supplementary Table 3 with fair or good quality status at timepoints 1–4 like in the DBN analysis. Clusters with common patterns were further grouped into four super-clusters through hierarchical clustering. The distribution of the patients in each super-cluster is presented in Table 1, where the average number of patients is 117 (27.72%), 108 (25.6%), 88 (20.85%), and 109 (25.83%) in super-clusters 1, 2, 3, and 4, respectively. Statistically significant differences were identified in the patient distribution for features “Hct”, “Lymph_abs_number”, “Lymph_percent”, “Neut_abs_number”, “Neut_percent”, “PO2_FiO2_ratio” regarding ICU admission and mortality. Additional differences among the patient subgroups were found in “AST” for ICU admission and in “ALP” and “LDH” for mortality. The detailed distribution of the ICU and non-ICU patients, as well as, the patients who died and those who survived per supercluster are presented in Supplementary Table 4.

3.3.2. *Trajectories*

With respect to the LGMM analysis, for all models the distribution

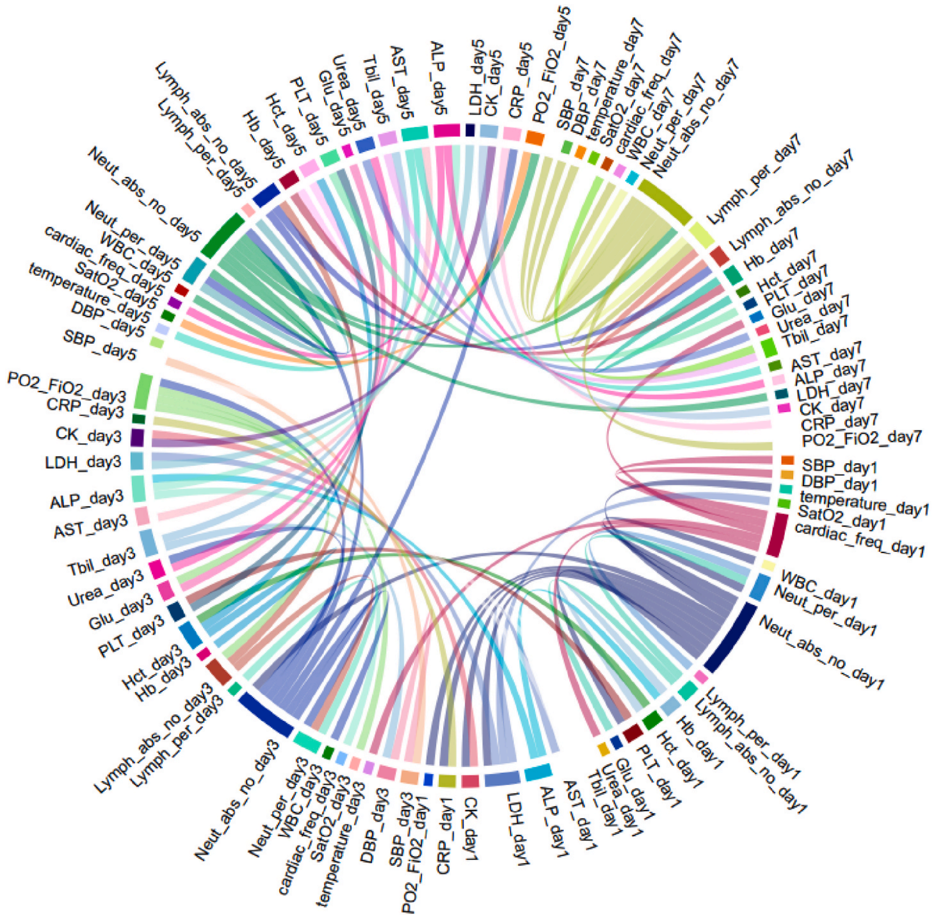

Fig. 3. A circular visualization of the DBN obtained in the present study based on the time series clinical data measured at 4 different time-points. This type of diagram presents the relationships (links) among the different nodes in our model. The more the connections among two nodes the stronger the relation among them.

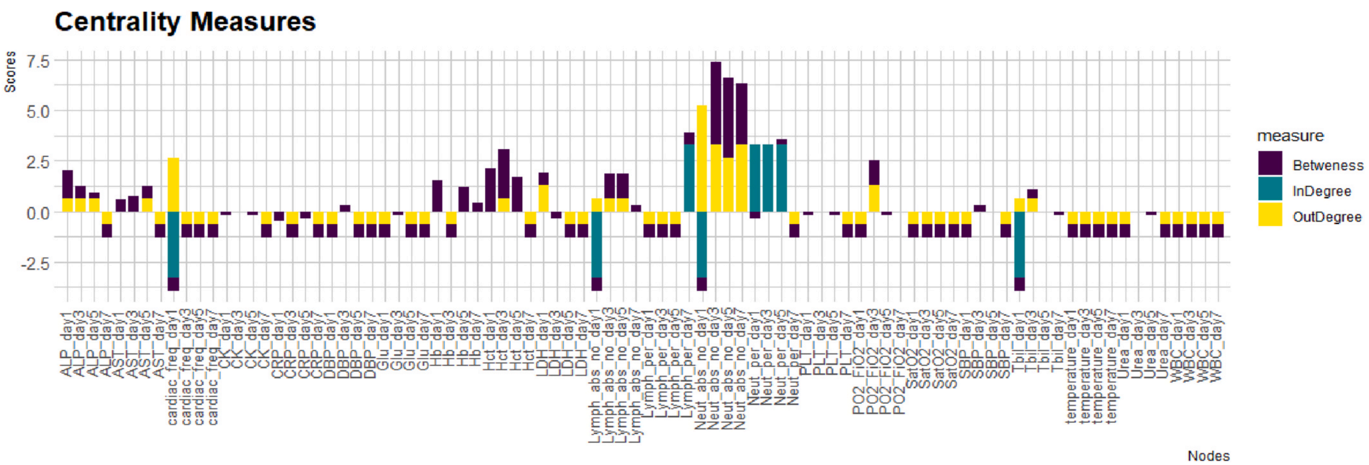

Fig. 4. The centrality measures extracted for each group of variables regarding the discrete time-points. The in- and out-degrees are shown for each node along with the betweenness centrality that provides the node's influence over information flow.

functions were approximated using either Beta or Splines transformations as they fitted better in terms of AIC than the linear transformations which deviated from normality. The fit statistics for the best clustering solutions per feature variable along with the size of clusters are provided in Supplementary Table 5, where the optimal clustering solutions resulted to either two or three clusters. According to Fig. 5, population trajectories were classified into two or three clusters. Each colored line represents the mean trajectory for a given cluster and the surrounding shaded area indicates the standard error of the mean. Additionally, a table with the absolute number of observations in the clusters is provided for each feature (Supplementary Table 5) along with statistical significances among the patient populations in the clusters based on ICU admission and mortality (Supplementary Table 6).

Overall, the LGMM analysis for "ALP", "AST", "Hct", "LDH", "Lymph_abs_number", "Lymph_percent" and "Neut_abs_number" resulted in 2-cluster solutions while for "cardiac_frequency", "Neut_percent" and

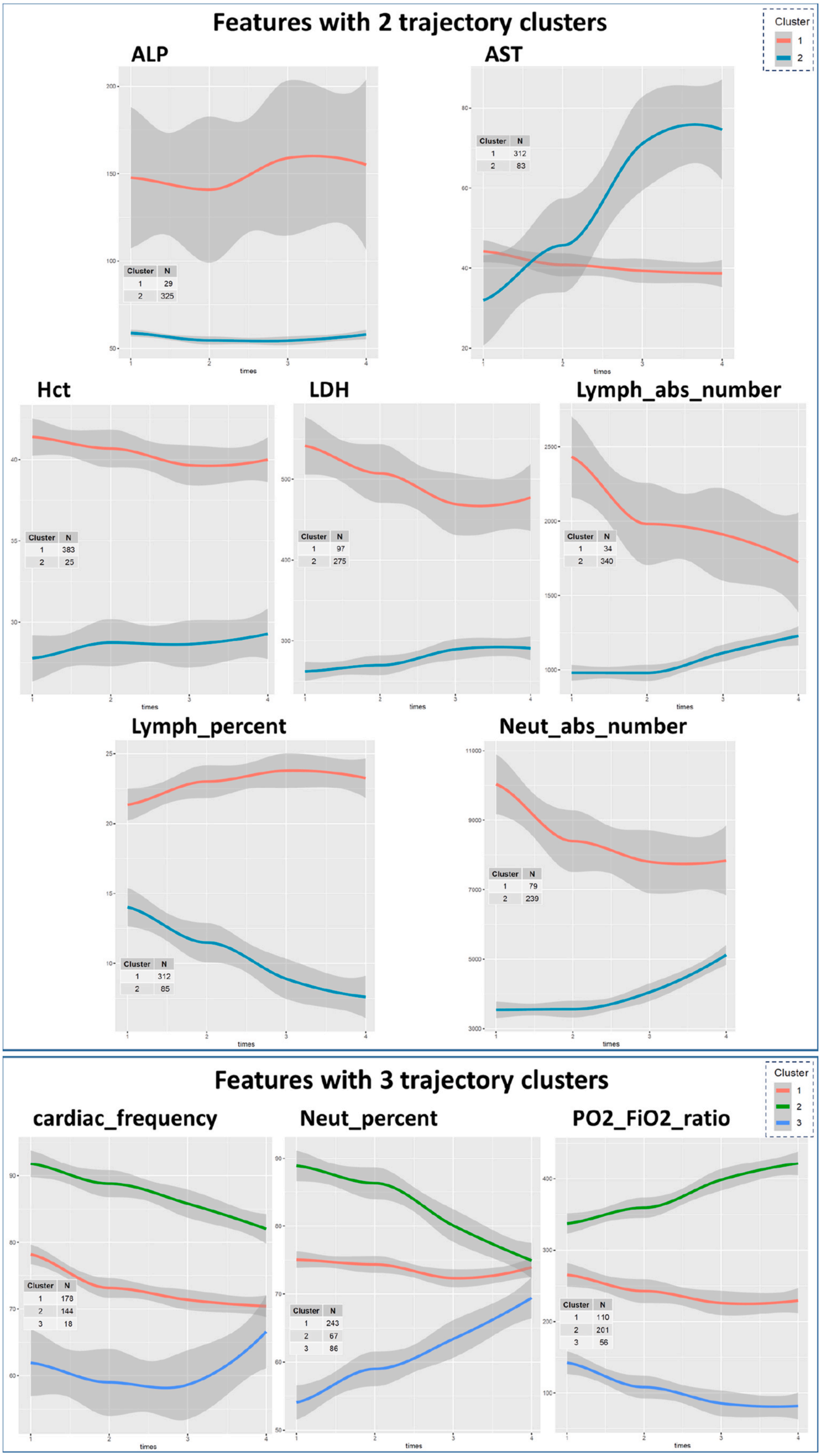

Fig. 5. The patterns of trajectory clusters identified for each feature.

Table 1
Number of patients assigned in each SOMs super-cluster for the most important features from the DBNs (p-values in bold denote statistically significant differences among the distributions of the ICU against the non-ICU patients and the patients who survived against those who died across the clusters).

Feature	Patient distribution in each super-cluster				p-value[a]	
	C1	C2	C3	C4	ICU	mortality
ALP	88	223	83	28	0.732	**0.04**
AST	173	71	92	86	**0.005**	0.285
cardiac_frequency	86	68	145	123	0.905	0.103
Hct	107	167	44	104	**<0.001**	**0.0001**
LDH	80	61	101	180	0.061	**0.005**
Lymph_abs_number	82	82	84	174	**0.015**	**0.033**
Lymph_percent	102	105	79	136	**0.024**	**0.0007**
Neut_abs_number	130	148	95	49	**0.016**	**0.005**
Neut_percent	148	95	74	105	**0.0008**	**0.0003**
PO2_FiO2_ratio	132	74	87	129	**<0.001**	**0.004**
Tbil	166	89	79	88	1	0.319
Average patient distribution	117	108	88	109		

[a] A Fisher's exact test was applied where the confidence level was set to 95%.

“PO2_FiO2”, it resulted in 3-cluster solutions. According to Supplementary Table 6, significant differences were identified in the patient distribution among the trajectory clusters for features “LDH”, “Lymph_percent” and “POS_FiO2_ratio”, regarding ICU admission and mortality. Additional differences were detected in “Hct”, “Neut_abs_number” and “cardiac_freq” for mortality.

3.4. *Classification performance for ICU admission and mortality*

Three case studies were investigated which involve the classification of patients for ICU admission and mortality (Table 2) based on: (i) the 51 time-series clinical data across the first 4 timepoints with and without the inclusion of the 32 features with the clustering labels from the SOMs, (ii) the 11 features from the DBNs analysis with and without the clustering labels from the SOMs, and (iii) only with the clustering labels from the SOMs. In case study 1, the contribution of the clustering labels from the SOMs enhanced the sensitivity by 1% and the specificity by 2% of the classifier for ICU admission against the use of the time-series data only. In case study 2, the contribution of the clustering labels from the SOMs enhanced the sensitivity and specificity of the classifier for ICU admission by 4% compared against the use of the best features from the DBNs, as well as, by 3% in sensitivity and 2% in specificity for mortality (Table 2). In case study 3, the use of the clustering labels from the SOMs yielded favorable classification performance. The performance evaluation results with and without the SOMs clustering labels for the best features from the DBNs are presented in Supplementary Table 7.

According to Table 2 and Supplementary Table 7, the performance of the classifiers was higher using the clustering labels from the SOMs for both mortality (in case study 1) and ICU admission (in case study 2), thus highlighting the positive impact of the DBNs and the SOMs during the training process. This can be also confirmed even in the case where no class imbalance handling is applied (Supplementary Table 8), where the performance of the classifiers remains higher using the clustering labels from the SOMs for both mortality (in case study 1) and ICU admission (in case study 3). Finally, the performance evaluation results with and without the clustering labels from the trajectories are depicted in Supplementary Table 9, where no performance improvement is observed.

The corresponding ROC curves are depicted in Fig. 6 for ICU and mortality classification across the three case studies from Table 2. Regarding the performance of the classifier for ICU admission, the average ROC was 0.89 for case 1, 0.91 for case 2, and 0.86 for case 3. As far as mortality classification is concerned, the average ROC was 0.83 for case 1, 0.76 for case 2, and 0.74 for case study 3.

3.5. *Risk factors for ICU admission and mortality*

According to Fig. 7, the risk factor analysis highlighted the following features as important (i.e., the top five features) for ICU admission in case study 1 (with the clustering labels from the SOMs): O2_supply_type_day5”, “O2_supply_type_SOM”, “SatO2_day7”, “tachypnea_day5”, and “SBP_day7”. The rest of the features include “temperature_day7”, “secondary_O2_supply_lit_SOM”, “PCO2_day3”, “K_day3”, and “DBP_day3”. Regarding mortality, the most informative features for decision making, include the: “Lymph_percent_day7”, “Urea_day5, “ALP_day1”, “Neut_percent_day7”, and “Hb_day1”. Additional features include the “tachypnea_day_3”, “INR_day1”, “PO2_FiO2_ratio_day5”, “hs_TPN_day1”, and “FiO2_day5”. The important features with the “SOM” tag denote the features with the clustering labels (Supplementary Table 3).

According to Fig. 8, the risk factor analysis indicated the following features as important for ICU admission in case study 2 (with the clustering labels from the SOMs): “PO2_FiO2_ratio_day5”, “Lymph_abs_number_day5”, “O2_supply_type_SOM”, “PO2_FiO2_ratio_day7”,

Table 2
Performance evaluation results from the GBT for ICU and mortality classification across different cases with donwsampling using the SOMs clustering labels from all the 32 continuous features (with blue color: specifications with the best or equal classification performance).

Case	Outcome	SOMs	accuracy	sensitivity	specificity	AUC
Case study 1*: 51 features across 4 timepoints with and without the clustering labels from the SOMs	**death**	no	0.74	0.74	0.76	0.83
	death	yes	0.74	0.74	0.76	0.83
	ICU	no	0.78	0.79	0.79	0.88
	ICU	yes	0.79	0.80	0.82	0.89
Case study 2*: 11 features from DBNs across 4 timepoints with and without the clustering labels from the SOMs	**death**	no	0.67	0.67	0.70	0.74
	death	yes	0.70	0.70	0.72	0.76
	ICU	no	0.78	0.79	0.78	0.87
	ICU	yes	0.83	0.83	0.82	0.91
Case study 3*: Only with the clustering labels from the SOMs	**death**	yes	0.67	0.67	0.68	0.74
	ICU	yes	0.80	0.80	0.82	0.86

*Random downsampling with replacement was applied to deal with the underlying class imbalance.

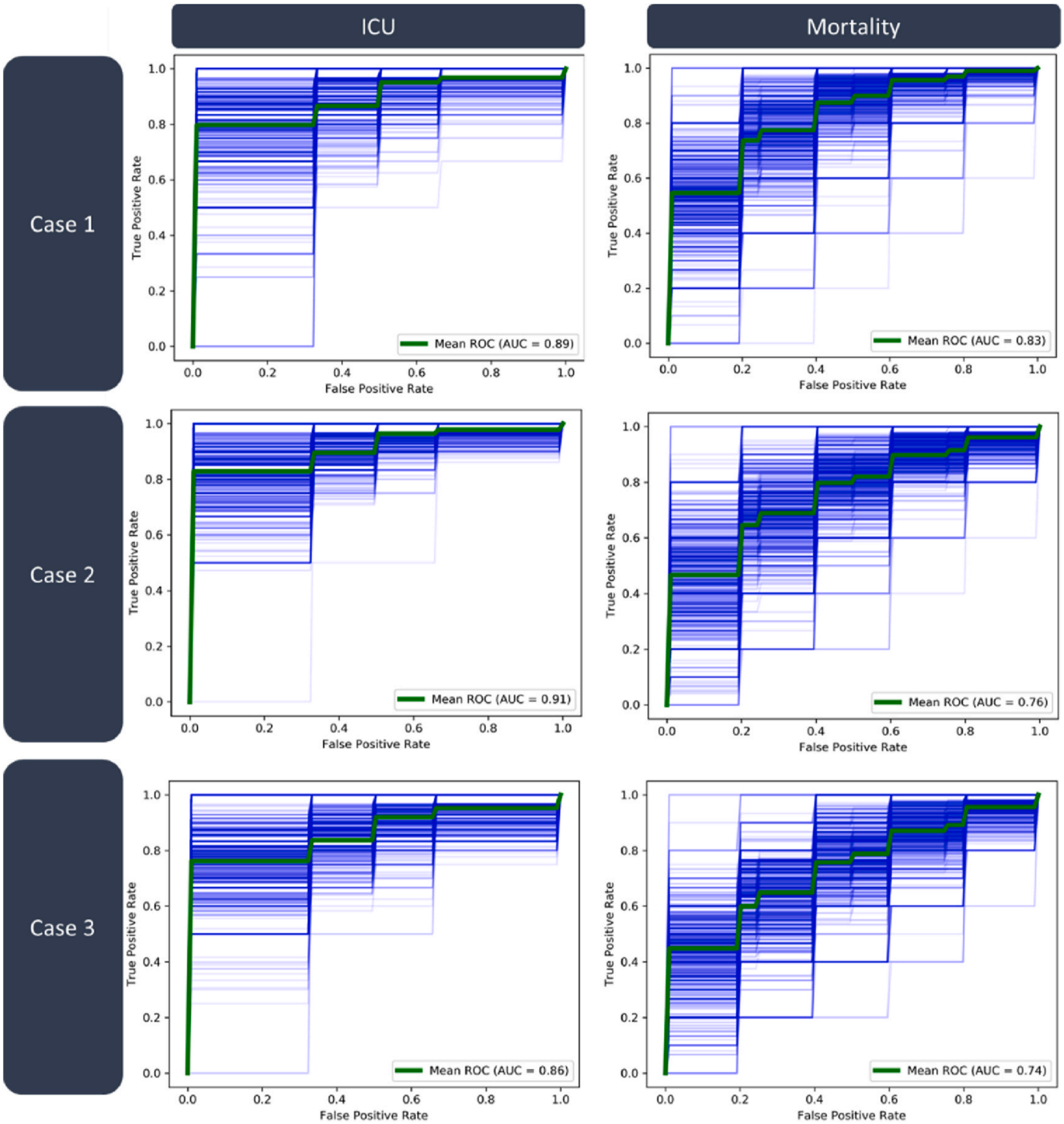

Fig. 6. Performance evaluation results for the GBT with the clustering labels from the SOMs. The line in bold denotes the average ROC across 100 iterations of the downsampling process.

and “Lymph_percent_day3”, among others. Regarding mortality, the most important features for decision making include the: “PO2_FiO2_ratio_day5”, “Hct_day1”, “ALP_day1”, “LDH_day5”, and “Neut_abs_number_day7”, among others.

According to Fig. 9, the analysis highlighted the following features as important for ICU admission in the case study 3: “O2_supply_type_SOM”, “temperature_SOM”, “secondary_O2_supply_lit_SOM”, “SatO2_SOM”, and “cardiac_frequency_SOM”, among others. Regarding mortality classification, the most important features include the: “SatO2_SOM”, “secondary_O2_supply_lit_SOM”, “Na_SOM”, “ALP_SOM, and “Creatinine_SOM”, among others.

In all cases, the clustering labels from the SOMs regarding the O2 supply type and the feature “ALP” were prominent for ICU admission and mortality, respectively (these features have been denoted with asterisks in Figs. 7–9).

3.6. *Inclusion of additional information (demographics, clinical data, treatments)*

An additional experiment was conducted to evaluate the contribution of baseline data (Supplementary Table 1) including demographics (e.g., age, gender, patient history), clinical (e.g., fever, fatigue, dyspnea), and treatments (e.g., administration of various therapeutic treatments, such as, statin, betablocker, corticosteroids) in the case study where the GBTs achieved the best performance in Table 2 (i.e., case study 2). According to Table 3, the inclusion of demographics, clinical, and treatments did not yield any improvement in the performance of the classifier for ICU admission. On the other hand, the sensitivity of the classifier for mortality was improved by 4% using the demographic data. The specificity was improved by 4% in the case where the demographics are included and by 1% in the case where the baseline clinical data and the treatments were included.

4. Discussion

We presented a straightforward workflow which combines DBNs with SOMs to derive homogeneous clusters of patients with COVID-19 based on a subset of features that have the highest degree and connectivity across multiple timepoints. The clustering labels from the SOMs were used to enrich the existing time-series clinical and laboratory data

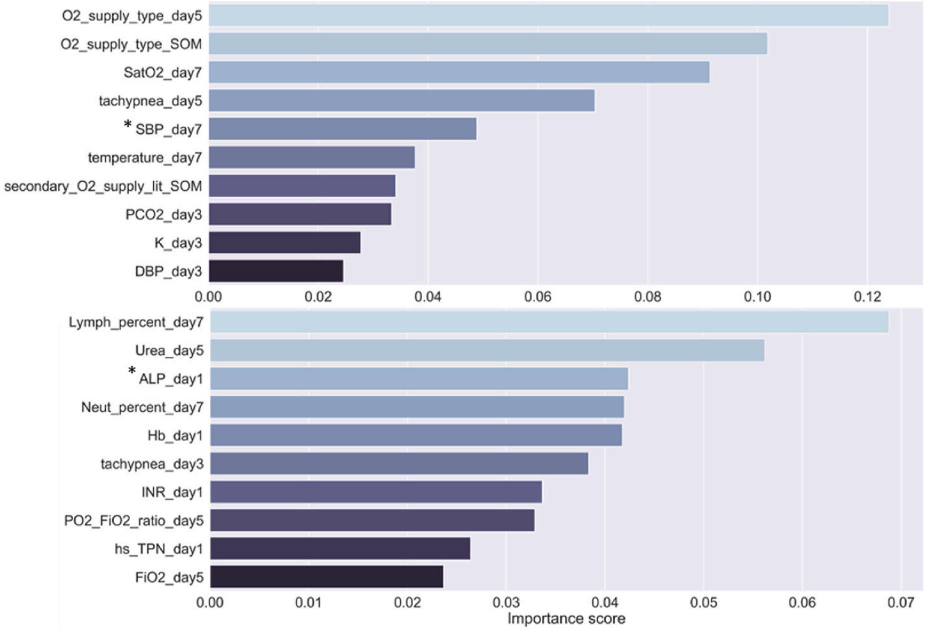

Fig. 7. Feature importance for ICU admission (on top) and mortality (on bottom) from case study 1 with the clustering labels from the SOMs.

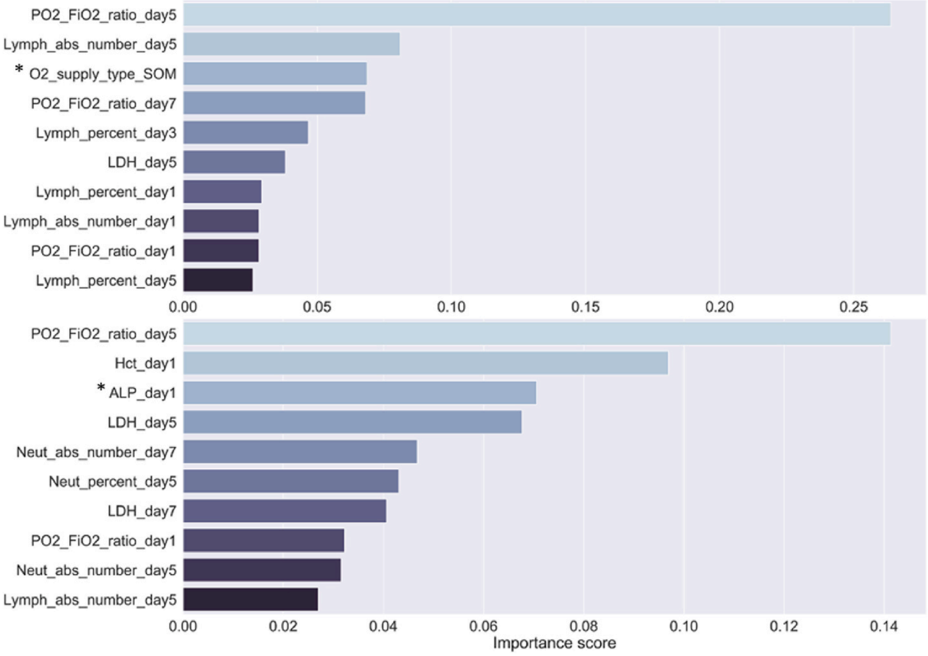

Fig. 8. Feature importance for ICU admission (on top) and mortality (on bottom) from case study 2 with the clustering labels from the SOMs.

with meta information yielding an increase in the performance of classification models for ICU admission and mortality. Our results highlight the contribution of the extracted patient subgroups from the SOMs along with the dynamically associated features with increased connectivity from the DBNs towards the improvement of the classification performance for ICU admission (sensitivity 0.83; specificity 0.83) and mortality (sensitivity 0.74; specificity 0.76). The number of lymphocytes, SatO2, PO2/FiO2, and O2 supply type were highlighted as prominent risk factors for ICU admission and the percentage of neutrophils and lymphocytes, PO2/FiO2, LDH, and ALP for mortality, among others.

Concerning the findings of the DBN model, two main variables were identified as significant predictors towards ICU admission and patients' mortality: cardiac frequency and neutrophil absolute number. As illustrated in Figs. 3 and 4 these two variables have the highest interconnectedness when they are measured at the baseline (i.e., day 1). Moreover, according to the DBN analysis, higher dependencies were yielded among these main contributors and the clinical variable of "PO2_FiO2". A more thorough investigation of these dependencies would permit the identification of major factors at the baseline or during the follow-up period that contribute to risk stratification of COVID-19 patients.

The presented DBN-based analysis provides the framework to extract causal and reasonable trajectories over time concerning the measurement of clinical and other related to COVID-19 data at discrete timepoints. Based on these findings we should mention that high degree of betweenness centrality has been observed among the Neutrophil

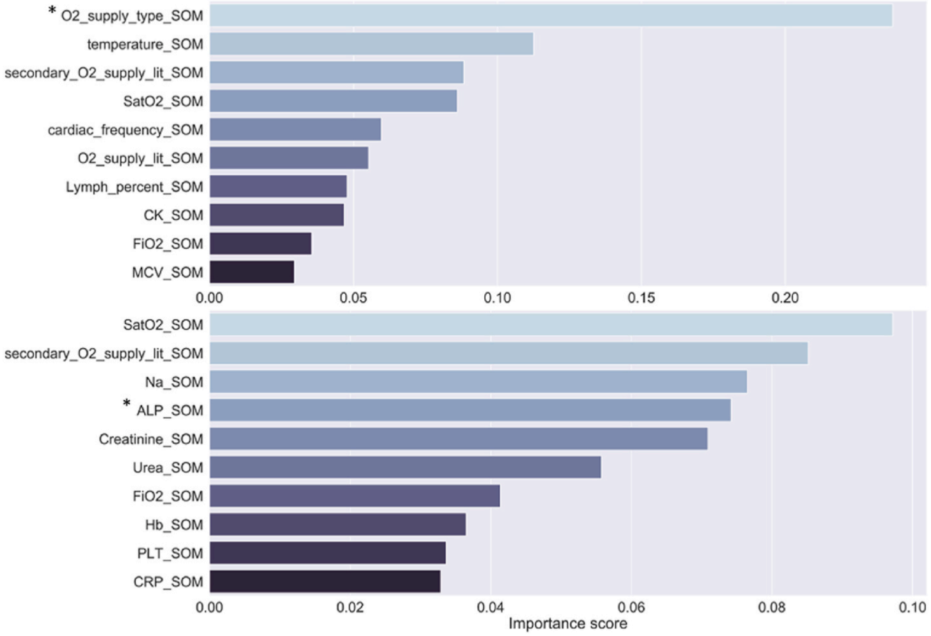

Fig. 9. Feature importance for ICU admission (on top) and mortality (on bottom) from case study 3 with the clustering labels from the SOMs.

Table 3
Performance evaluation results for case study 2 before and after the inclusion of demographics, clinical data and treatments (with blue color: specifications with the best or equal classification performance).

Outcome	Accuracy	Sensitivity	Specificity	AUC
ICU admission				
Before	0.83	0.83	0.82	0.91
After adding demographic data	0.83	0.83	0.81	0.90
After adding clinical data	0.81	0.81	0.80	0.89
After adding treatments	0.83	0.83	0.82	0.91
Mortality				
Before	0.70	0.70	0.72	0.76
With demographic data	0.74	0.74	0.76	0.82
With clinical data	0.70	0.70	0.73	0.76
With treatments	0.70	0.70	0.73	0.76

number (“Neut_abs_no_day1”) and the cardiac frequency (“cardiac_freq”) at day 1 variables. The analysis of clinical data measured at discrete time-points through DBN modeling would permit the identification of new dependencies among significant factors regarding the ICU admission of COVID-19 patients. In addition, the detection of new dependencies and relationships between clinical-oriented variables that characterize the progression of the disease would allow better decision making in the clinical management of the disease as well as the suggestion of targeted therapy that could decrease the mortality rates.

Significant differences were identified in the patient distribution across the four super-clusters from the SOMs analysis and particularly for the features “Hct”, “Lymph_abs_number”, “Lymph_percent”, “Neut_abs_number”, “Neut_percent” and “PO2_FiO2_ratio”, regarding ICU admission and mortality. Additional significant differences were detected in “AST” for ICU admission and in “ALP” and “LDH” for mortality. These findings are in line with those obtained by the trajectories analysis (Supplementary Tables 5 and 6), where statistically significant differences were identified among the patient distribution in the clusters for the features “LDH”, “Lymph_percent” and “POS_FiO2_ratio”, regarding ICU admission and mortality. Additional statistically significant differences were detected in “Hct”, “Neut_abs_number” and “cardiac_freq” for mortality.

The most important features, as denoted by the DBNs, were utilized in the SOMs to extract homogeneous clusters of COVID-19 patients with common clinical profiles. Subsequently, MARTs were trained on the aggregated features from the DBNs and the new features from the SOMs yielding robust classifiers for ICU admission and mortality with an increase by 1% in sensitivity and 2% in specificity for ICU admission in case study 1, as well as an increase by 4% in sensitivity and specificity for ICU admission and by 3% in sensitivity and 2% in specificity in case study 2, compared to the classifiers trained with the clustering labels from the SOMs. The contribution of demographics-related data yielded an increase in accuracy by 4% and in AUC by 6% for mortality (Table 3) which suggests that age has a high impact on mortality in hospitalized COVID-19 patients.

The number of lymphocytes, SatO2, PO2/FiO2 and O2 supply type were highlighted as major risk factors for ICU admission and the percentage of neutrophils and lymphocytes, PO2/FiO2, LDH, and ALP for mortality. According to Table 4, our findings are in line with related risk factors that were reported in the literature. More specifically, the neutrophils infiltration has been found to drive necroinflammation during coronavirus in Ref. [38] whereas in Ref. [39] the neutrophil to lymphocyte ratio has been highlighted as a risk factor for the severity of COVID-19. Additional risk factors for mortality include the “Hb” which has been highlighted also in Ref. [40] as an independent risk factor for the mortality in COVID-19 patients and the “INR” which has been linked with COVID-19 severity in Ref. [41]. The prognostic value of troponin elevation has been identified in Ref. [42] and particularly in patients with underlying cardiovascular diseases. The “PO2_FiO2_ratio” along with the “FiO2” have been identified as independent risk factors for in-hospital mortality in patients with COVID-19 [43]. Likewise, LDH has been found as an independent risk factor of severe COVID-19 in Ref. [44] while tachypnea and low SBP have been strongly associated with in-hospital mortality in COVID-19 [45]. Additional risk factors for ICU admission include the supply oxygen type which is highly associated with COVID-19 severity, and SatO2 which serves as a predictor of mortality in adult patients with COVID-19 [46]. The relationship between mortality and ALP has also been demonstrated in Refs. [47,48] which underline the clinical need for further investigation of elevated serum alkaline phosphatase levels as a mechanism of liver injury in COVID-19. In addition, this study goes beyond the state of the art by combining DBNs with SOMs and trajectories to derive homogeneous clusters of patients with COVID-19 based on a subset of features that have the highest degree and connectivity across multiple timepoints.

Unlike the existing studies (Table 4) which focus on the direct application of machine learning algorithms for the development of ICU admission and mortality classifiers and the detection of related risk factors, the proposed approach places particular emphasis on the dynamic modeling of features across multiple time-points to extract the most informative ones. The latter are utilized to derive homogeneous clusters of COVID-19 patients with similar clinical profiles based on the SOMs and the trajectory analysis. The extracted clustering information is then combined with the input data to enhance the robustness of the classifiers for ICU admission and mortality.

Table 4
Comparison with the state-of-the-art studies for ICU admission and mortality in COVID-19.

Study	Method	Risk factors
Subudhi et al., 2021 [12]	Ensemble-based algorithms to predict ICU admission and mortality across 3597 COVID-19 patients.	Risk factors: CRP, LDH, O2 saturation for ICU admission and neutrophil and lymphocytes for mortality.
Cheng et al., 2020 [10]	Random forests for risk stratification based on time-series data across 1987 unique patients diagnosed with COVID-19.	A risk prioritization tool that predicts the need for ICU admission within 24h to optimize the flow of operations within the hospitals.
Dan et al., 2020 [13]	Ensemble learning to objectively identify an optimal combination of factors that predicts ICU admissions across 733 COVID-19 patients.	The number of lymphocytes was involved in all prediction tasks with the highest AUC score.
Aznar et al., 2021 [15]	Multipurpose algorithms (boosting ensembles, artificial neural networks) to estimate the risk of ICU admission or mortality among 3623 patients with COVID-19.	The final model achieved good discrimination for the external validation set (AUC 0.821). A cut-off of 0.4 yields sensitivity and specificity 0.71 and 0.78, respectively.
Fernades et al., 2021 [16]	Predict the risk for COVID-19 severity by training multipurpose algorithms across 3280 patients.	High predictive performance (average ROC 0.92) with the following risk factors: lymphocytes, C-reactive protein, and Braden Scale.
Guan et al., 2021 [14]	GBTs were trained on 1270 COVID-19 patients from Wuhan to detect risk factors.	Age, CRP, and LDH were identified as prominent features for COVID-19 mortality.
Chen et al., 2021 [11]	Bagging methods were applied on clinical data from 362 patients with confirmed COVID-19.	Age, hypertension, gender, diabetes, absolute neutrophil count, IL-6, and LDH were identified as risk factors for COVID-19 severity.
Current work	DBNs combined with SOMs to derive homogeneous clusters of patients with COVID-19 which were used to enrich the existing time-series clinical and laboratory data with meta information to increase the performance of classification models for ICU admission and mortality.	Risk factors: number of lymphocytes, SatO2, PO2/FiO2, and O2 supply type as risk factors for ICU admission and the percentage of neutrophils and lymphocytes, PO2/FiO2, LDH, and ALP for mortality. Classification performance for ICU admission with sensitivity: 0.83 and specificity: 0.83 (AUC 0.91), and mortality with sensitivity: 0.74 and specificity: 0.76 (AUC 0.83).

5. Conclusions and future work

In this work, we used DBN modeling to predict probable and reasonable trajectories over time, considering the measurement of clinical data in discrete time-points. In addition, we identified underlying probabilistic relationships among prominent risk factors of COVID-19 for both ICU admission and mortality. The clustering and trajectory analysis revealed major factors, including the number of lymphocytes, PO2/FiO2, percentage of neutrophils and lymphocytes, LDH, and ALP at the baseline or during the follow-up as prominent for ICU admission and mortality in COVID-19. The contribution of the extracted clusters, the trajectories and the dynamically associated clinical data yielded an improved classification performance both for ICU admission (sensitivity 0.83, specificity 0.83) and mortality (sensitivity 0.74, specificity 0.76). Thorough investigation of the derived patient subgroups (i.e., clusters and trajectories) would permit the identification of major factors at the baseline or during the follow-up period that contribute to risk stratification of COVID-19 patients. The sensitivity of the classifier for mortality was improved by 4% using demographic-related data while the specificity was improved by 4% in the case where the baseline clinical data are included and by 3% in the case where the demographics and the therapies-related data were incorporated. The features “number of lymphocytes”, “SatO2”, “PO2/FiO2” and “O2 supply type” were highlighted as risk factors for ICU admission and the percentage of neutrophils and lymphocytes, PO2/FiO2, LDH and ALP for mortality, among others. Although most of the existing studies (Table 4) focus on the development of ICU admission and mortality classifiers without taking into consideration the underlying dynamic associations among the data, the proposed method combines dynamic modeling with clustering analysis to identify subgroups of COVID-19 patients with common clinical profiles which are in turn utilized for the development of robust classifiers for ICU admission and mortality.

As a future work, we plan to extend the population size and further enrich the clinical data to enhance the performance of the classifiers for ICU admission and mortality, as well as, to capture dynamic associations among different phenotypes of COVID-19 across additional timepoints to better understand the underlying pathogenic mechanisms of the disease based on deep learning methods.

Declaration of competing interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Acknowledgements

The research work was supported by the Hellenic Foundation for Research and Innovation (HFRI) under the HFRI PhD Fellowship grant (Fellowship Number: 1357).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2021.105176.

References

[1] World Health Organization, WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020, Available from: https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19—11-march-2020. (Accessed 14 March 2020).

[2] D.P. Oran, E.J. Topol, The proportion of SARS-CoV-2 infections that are asymptomatic: a systematic review, Ann. Intern. Med. 174 (5) (2021) 655–662.

[3] Z. Wu, J.M. McGoogan, Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: summary of a report of 72314 cases from the Chinese center for disease control and prevention, JAMA 323 (13) (2020) 1239–1242.

[4] X. Yang, X. Yang, Y. Yu, J. Xu, H. Shu, H. Liu, et al., Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study, Lancet Respir. Med. 8 (5) (2020) 475–481. Lancet Respir Med, 2020. 8(5): pp. 475–481.

[5] C.M. Petrilli, S.A. Jones, J. Yang, H. Rajagopalan, L. O'Donnell, Y. Chernyak, L. I. Horwitz, Factors associated with hospital admission and critical illness among 5279 people with coronavirus disease 2019 in New York City: prospective cohort study, BMJ 369 (2020).

[6] E.K. Stokes, L.D. Zambrano, K.N. Anderson, E.P. Marder, K.M. Raz, S.E.B. Felix, K. E. Fullerton, Coronavirus disease 2019 case surveillance—United States, January 22–May 30, 2020, MMWR (Morb. Mortal. Wkly. Rep.) 69 (24) (2020) 759.

[7] S. Richardson, J.S. Hirsch, M. Narasimhan, J.M. Crawford, T. McGinn, K. W. Davidson, Northwell COVID-19 Research Consortium, Presenting characteristics, comorbidities, and outcomes among 5700 patients hospitalized with COVID-19 in the New York City area, JAMA 323 (20) (2020) 2052–2059.

[8] A.B. Docherty, E.M. Harrison, C.A. Green, H.E. Hardwick, R. Pius, L. Norman, M. G. Semple, Features of 20 133 UK patients in hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: prospective observational cohort study, BMJ (2020) 369.

[9] D.S.W. Ting, L. Carin, V. Dzau, T.Y. Wong, Digital technology and COVID-19, Nat. Med. 26 (4) (2020) 459–461.

[10] F.Y. Cheng, H. Joshi, P. Tandon, R. Freeman, D.L. Reich, M. Mazumdar, A. Kia, Using machine learning to predict ICU transfer in hospitalized COVID-19 patients, J. Clin. Med. 9 (6) (2020) 1668.

[11] Y. Chen, L. Ouyang, F.S. Bao, Q. Li, L. Han, H. Zhang, S. Chen, A Multimodality Machine Learning Approach to Differentiate Severe and Nonsevere COVID-19: Model Development and Validation, J. Med. Internet Res. 23 (4) (2021).

[12] S. Subudhi, A. Verma, A.B. Patel, C.C. Hardin, M.J. Khandekar, H. Lee, R.K. Jain, Comparing machine learning algorithms for predicting ICU admission and mortality in COVID-19, NPJ digital medicine 4 (1) (2021) 1–7.

[13] T. Dan, Y. Li, Z. Zhu, X. Chen, W. Quan, Y. Hu, H. Cai, Machine learning to predict ICU admission, ICU mortality and survivors' length of stay among COVID-19 patients: toward optimal allocation of ICU resources, in: 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2020, December, pp. 555–561.

[14] X. Guan, B. Zhang, M. Fu, M. Li, X. Yuan, Y. Zhu, Y. Lu, Clinical and inflammatory features based machine learning model for fatal risk prediction of hospitalized COVID-19 patients: results from a retrospective cohort study, Ann. Med. 53 (1) (2021) 257–266.

[15] R. Aznar-Gimeno, L.M. Esteban, G. Labata-Lezaun, R. del-Hoyo-Alonso, D. Abadia-Gallego, J.R. Paño-Pardo, M. Serrano, A clinical decision web to predict ICU admission or death for patients hospitalised with COVID-19 using machine learning algorithms, Int. J. Environ. Res. Publ. Health 18 (16) (2021) 8677.

[16] F.T. Fernandes, T.A. de Oliveira, C.E. Teixeira, A.F. de Moraes Batista, G. Dalla Costa, A.D.P. Chiavegatto Filho, A multipurpose machine learning approach to predict COVID-19 negative prognosis in São Paulo, Brazil, Sci. Rep. 11 (1) (2021) 1–7.

[17] V.C. Pezoulas, K.D. Kourou, F. Kalatzis, T.P. Exarchos, A. Venetsanopoulou, E. Zampeli, D.I. Fotiadis, Medical data quality assessment: on the development of an automated framework for medical data curation, Comput. Biol. Med. 107 (2019) 270–283.

[18] P. Arora, D. Boyne, J.J. Slater, A. Gupta, D.R. Brenner, M.J. Druzdzel, Bayesian networks for risk prediction using real-world data: a tool for precision medicine, Value Health 22 (4) (2019) 439–445.

[19] D. Heckerman, C. Meek, G. Cooper, A Bayesian approach to causal discovery, Computation, causation, and discovery 19 (1999) 141–166.

[20] S.Y. Kim, S. Imoto, S. Miyano, Inferring gene networks from time series microarray data using dynamic Bayesian networks, Briefings Bioinf. 4 (3) (2003) 228–235.

[21] B. Baur, S. Bozdag, A canonical correlation analysis-based dynamic bayesian network prior to infer gene regulatory networks from multiple types of biological data, J. Comput. Biol. 22 (4) (2015) 289–299.

[22] K. Murphy, S. Mian, Modelling Gene Expression Data Using Dynamic Bayesian Networks, vol. 104, Computer Science Division, University of California, Berkeley, CA, 1999. Technical report.

[23] A. Franzin, F. Sambo, B. Di Camillo, bnstruct: an R package for Bayesian Network structure learning in the presence of missing data, Bioinformatics 33 (8) (2017) 1250–1252.

[24] C. Proust-Lima, V. Philipps, B. Liquet, Estimation of Extended Mixed Models Using Latent Classes and Latent Processes: the R Package Lcmm, 2015 arXiv preprint arXiv:1503.00890.

[25] P. Melin, J.C. Monica, D. Sanchez, O. Castillo, Analysis of spatial spread relationships of coronavirus (COVID-19) pandemic in the world using self organizing maps, Chaos, Solit. Fractals 138 (2020) 109917.

[26] V.C. Pezoulas, C. Papaloukas, M. Veyssiere, A. Goules, A.G. Tzioufas, V. Soumelis, D.I. Fotiadis, A computational workflow for the detection of candidate diagnostic biomarkers of Kawasaki disease using time-series gene expression data, Comput. Struct. Biotechnol. J. 19 (2021) 3058–3068.

[27] D.R. de Lima Cabral, R.S.M. de Barros, Concept drift detection based on Fisher's Exact test, Inf. Sci. 442 (2018) 220–234.

[28] J. Boelaert, L. Bendhaiba, M. Olteanu, N. Villa-Vialaneix, SOMbrero: an r package for numeric and non-numeric self-organizing maps, in: Advances in Self-Organizing Maps and Learning Vector Quantization, Springer, Cham, 2014, pp. 219–228.

[29] S. Frankfurt, P. Frazier, M. Syed, K.R. Jung, Using group-based trajectory and growth mixture modeling to identify classes of change trajectories, Counsel. Psychol. 44 (5) (2016) 622–660.

[30] N. Ram, K.J. Grimm, Methods and measures: growth mixture modeling: a method for identifying differences in longitudinal change among unobserved groups, Int. J. Behav. Dev. 33 (6) (2009) 565–576.

[31] K.L. Nylund, T. Asparouhov, B.O. Muthén, Deciding on the number of classes in latent class analysis and growth mixture modeling: a Monte Carlo simulation study, Struct. Equ. Model.: A Multidiscip. J. 14 (4) (2007) 535–569.

[32] V. Ramaswamy, W.S. DeSarbo, D.J. Reibstein, W.T. Robinson, An empirical pooling approach for estimating marketing mix elasticities with PIMS data, Market. Sci. 12 (1) (1993) 103–124.

[33] T. Asparouhov, B. Muthén, Auxiliary variables in mixture modeling: three-step approaches using M plus, Struct. Equ. Model.: A Multidiscip. J. 21 (3) (2014) 329–341.

[34] G. van der Nest, V.L. Passos, M.J. Candel, G.J. van Breukelen, An overview of mixture modelling for latent evolutions in longitudinal data: Modelling approaches, fit statistics and software, Adv. Life Course Res. 43 (2020) 100323.

[35] C. Proust-Lima, V. Philipps, B. Liquet, Estimation of Extended Mixed Models Using Latent Classes and Latent Processes: the R Package Lcmm, 2015 arXiv preprint arXiv:1503.00890.

[36] T. Chen, C. Guestrin, August). GBT: a scalable tree boosting system, in: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.

[37] V.C. Pezoulas, K.D. Kourou, F. Kalatzis, T.P. Exarchos, E. Zampeli, S. Gandolfo, D. I. Fotiadis, Overcoming the barriers that obscure the interlinking and analysis of clinical data through harmonization and incremental learning, IEEE Open Journal of Engineering in Medicine and Biology 1 (2020) 83–90.

[38] B. Tomar, H.J. Anders, J. Desai, S.R. Mulay, Neutrophils and neutrophil extracellular traps drive necroinflammation in COVID-19, Cells 9 (6) (2020) 1383.

[39] S. Wang, L. Fu, K. Huang, J. Han, R. Zhang, Z. Fu, Neutrophil-to-lymphocyte ratio on admission is an independent risk factor for the severity and mortality in patients with coronavirus disease 2019, J. Infect. 82 (2) (2021) e16–e18.

[40] Z. Wang, Z. Du, F. Zhu, Glycosylated hemoglobin is associated with systemic inflammation, hypercoagulability, and prognosis of COVID-19 patients, Diabetes Res. Clin. Pract. 164 (2020) 108214.

[41] A. Zinellu, P. Paliogiannis, C. Carru, A.A. Mangoni, INR and COVID-19 severity and mortality: a systematic review with meta-analysis and meta-regression, Adv. Med. Sci. 66 (2) (2021) 372–380.

[42] E.M. Cordeanu, N. Duthil, F. Severac, H. Lambach, J. Tousch, L. Jambert, D. Stephan, Prognostic value of troponin elevation in COVID-19 hospitalized patients, J. Clin. Med. 9 (12) (2020) 4078.

[43] P. Santus, D. Radovanovic, L. Saderi, P. Marino, C. Cogliati, G. De Filippis, G. Sotgiu, Severity of respiratory failure at admission and in-hospital mortality in patients with COVID-19: a prospective observational multicentre study, BMJ Open 10 (10) (2020), e043651.

[44] Y. Han, H. Zhang, S. Mu, W. Wei, C. Jin, C. Tong, G. Gu, Lactate dehydrogenase, an independent risk factor of severe COVID-19 patients: a retrospective and observational study, Aging (N Y) 12 (12) (2020) 11245.

[45] T. Mikami, H. Miyashita, T. Yamada, M. Harrington, D. Steinberg, A. Dunn, E. Siau, Risk factors for mortality in patients with COVID-19 in New York City, J. Gen. Intern. Med. 36 (1) (2021) 17–26.

[46] F. Mejía, C. Medina, E. Cornejo, E. Morello, S. Vásquez, J. Alave, G. Málaga, Oxygen saturation as a predictor of mortality in hospitalized adult patients with COVID-19 in a public hospital in Lima, Peru, PLoS One 15 (12) (2020), e0244171.

[47] J.W. Goodall, T.A. Reed, M. Ardissino, P. Bassett, A.M. Whittington, D.L. Cohen, N. Vaid, Risk Factors for Severe Disease in Patients Admitted with COVID-19 to a Hospital in London, England: a Retrospective Cohort Study, vol. 148, Epidemiology & Infection, 2020.

[48] P. Sharma, A. Kumar, Metabolic dysfunction associated fatty liver disease increases risk of severe Covid-19, Diabetes & Metabolic Syndrome: Clin. Res. Rev. 14 (5) (2020) 825–827.