Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Multi-objective optimization determines when, which and how to fuse deep networks: An application to predict COVID-19 outcomes

Valerio Guarrasi\textsuperscript{a,b}, Paolo Soda\textsuperscript{a,c,*}

\textsuperscript{a} Unit of Computer Systems and Bioinformatics, Department of Engineering, University Campus Bio-Medico of Rome, Italy
\textsuperscript{b} Department of Computer, Control, and Management Engineering, Sapienza University of Rome, Italy
\textsuperscript{c} Department of Radiation Sciences, Radiation Physics, Biomedical Engineering, Umeå, University, Umeå, Sweden

\textbf{A R T I C L E I N F O}

\textbf{Keywords:}
- Deep-learning
- Multimodal learning
- Optimization

\textbf{A B S T R A C T}

The COVID-19 pandemic has caused millions of cases and deaths and the AI-related scientific community, after being involved with detecting COVID-19 signs in medical images, has been now directing the efforts towards the development of methods that can predict the progression of the disease. This task is multimodal by its very nature and, recently, baseline results achieved on the publicly available AlforCOVID dataset have shown that chest X-ray scans and clinical information are useful to identify patients at risk of severe outcomes. While deep learning has shown superior performance in several medical fields, in most of the cases it considers unimodal data only. In this respect, when, which and how to fuse the different modalities is an open challenge in multimodal deep learning. To cope with these three questions here we present a novel approach optimizing the setup of a multimodal end-to-end model. It exploits Pareto multi-objective optimization working with a performance metric and the diversity score of multiple candidate unimodal neural networks to be fused. We test our method on the AlforCOVID dataset, attaining state-of-the-art results, not only outperforming the baseline performance but also being robust to external validation. Moreover, exploiting XAI algorithms we figure out a hierarchy among the modalities and we extract the features' intra-modality importance, enriching the trust on the predictions made by the model.

1. Introduction

It has been two years since the severe acute respiratory syndrome COVID-19 has struck the world with a pandemic causing millions of cases and deaths. Since the beginning, the scientific community has tried to contain the spread and the number of victims by studying the virus from multiple perspectives and looking for the best cures.

Among the various initiatives to combat the rising of the pandemic, researchers, practitioners and enterprises introduced new artificial intelligence (AI) methods and tools to process chest X-ray (CXR) and computed tomography (CT) examinations to replace or to supplement the reverse transcriptase-polymerase chain reaction (RT-PCR) tests. Both medical imaging procedures are instrumental in helping radiologists determine the source of symptoms, stratify the disease severity, and establish the best treatment plan for the patient’s specific needs. On the other hand, CXR helps indicate abnormal formations of a large variety of chest diseases by using a very small amount of radiation. Reticular alteration (up to 5 days from the symptoms’ onset) and ground-glass opacity (following 5 days from the symptoms’ onset) are the CXR most frequent lesions in COVID-19 positive patients. The patients’ consolidation gradually increases over time, striking mainly bilateral, peripheral, middle/lower locations. On the other hand, CT delivers a much higher level of detail, offering a computerized 360-degree view of the lungs’ structures. However, by choosing this modality, we observed the lack of availability of machines’ slots, the difficulty of moving bedridden patients, and long sanitation times. On the contrary, X-ray equipment is much smaller, less complex, and with lower costs than CT scans. For these reasons, in this work we focus on the use of CXR scans, that fit well with the COVID-19 crisis.

Most of the AI models for medical image-based applications, in particular those exploiting deep-learning (DL) models, consider only pixel-value data neglecting information available in other modalities. However, interpreting imaging findings is multimodal by its very nature and, hence, AI needs to be able to process together data coming from various modalities [1]. Indeed, while multimodal learning with traditional approaches is a well-established field of study [1,2], the potential of deep-learning shown processing unimodal data has recently motivated the rise of multimodal deep-learning (MDL), which aims to treat multimodal data by using deep network-based approaches. This

\textsuperscript{*} Corresponding author at: Department of Radiation Sciences, Radiation Physics, Biomedical Engineering, Umeå, University, Umeå, Sweden.

\textit{E-mail addresses:} valerio.guarrasi@unicampus.it (V. Guarrasi), paolo.soda@umu.se (P. Soda).

https://doi.org/10.1016/j.compbiomed.2023.106625

Received 28 October 2022; Received in revised form 18 January 2023; Accepted 28 January 2023

Available online 2 February 2023

0010-4825/© 2023 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).
should be very useful to progress in understanding the world around us, biomedicine included [3]. In the specific context of the COVID-19 pandemic, MDL can support the patients’ stratification by mining together images and clinical information [4].

The literature agrees that the three main open questions in MDL are when, which and how the modalities must be joined to exploit their potential [5]. Usually, MDL models are constructed by finding or combining the best model architecture for each modality, which are then combined by researchers and practitioners on the basis of the nature of the data, the task at hand and the networks’ structure. However, this handcrafted approach does not necessarily provide the best ensemble [6]. To overcome these limitations, here we present a novel joint fusion technique that algorithmically finds the optimal set of deep architectures to be combined, whatever the modality they belong to, by maximizing both a classification evaluation metric and a measure of diversity among the learners. Indeed, diversity is an estimate of how the networks interpret data differently [7]. Next, the method trains the resulting architecture in an end-to-end fashion to carry out the task at hand. Further to determine which models and modalities should be fused, this work investigates when to join the different modalities and how to embed them in powerful multimodal data representations. We apply the method to stratify patients suffering from the COVID-19 syndrome by predicting the severity of the outcome using both clinical and CXR scans, showing that we achieve state-of-the-art results. Furthermore, to improve trust and transparency of our MDL approach, we present a novel combination of Explainable Artificial Intelligence (XAI) techniques not only illustrating the reasoning behind the decisions taken by the model, but also showing the relative contribution of each modality in making the decision. This XAI approach, as well as the robustness showed by our proposal on an external validation set, boosts reliability and trust in our overall results, thus addressing the main issue to obtain trust in AI technologies, as discussed in [8].

With respect to the previous work [4], this contribution presents three main novelties. First, we introduce a novel method exploiting multi-objective optimization to find the optimal fusion of multimodal models, which are selected from a large pool of many. Indeed, such previous contribution offered a baseline fusion paradigm where the multimodal features are exploited by early fusion or by baseline joint learning working with two fixed network architectures for CXR images and clinical data. In particular, the proposed optimized multimodal fusion exploits an end-to-end framework, relying on a combination of an evaluation and a diversity score. Second, our framework goes hand-in-hand with XAI algorithms which have not yet been investigated in this application, which permits to discover the modalities’ hierarchy and features’ intra-modality importance [9]. Third, this contribution not only performs experiments according to the procedure presented in [4], but it also works with other 283 new patients now available within the AIforCOIVD repository that we used as an external validation set.

The manuscript is organized as follows. Section 2 introduces the state-of-the-art of both multimodal fusion methods and DL applications of COVID-19 stratification. Section 3, illustrates the dataset on which the experiments were performed. Section 4 describes the novel fusion method presented in this work and the XAI algorithms used for interpretation. Section 5 introduces the experimental configuration in the context of COVID-19 prognosis prediction, whilst the corresponding results and explanations are offered in Section 6. Finally, Section 7 provides concluding remarks.

2. Background

With the spread of the epidemic, there has been a growth of interest in employing novel AI methods to fight the pandemic [10]. The vast literature ranges from the use of traditional statistical models to the more complex DL networks [11]. As it is out of the scope of this manuscript to mention all the literature, here we give an overview of the state-of-the-art of DL applications predicting the disease’s outcome, focusing on those that used supervised MDL models and exploited a joint fusion between medical imaging and clinical data. Since the MDL method presented in this manuscript fits well with XAI algorithms, we also give an overview of this emerging field of study.

2.1. Multimodal deep learning

Multimodal data is composed of data originated by different sources observing a phenomena. The objective of MDL is to work with the data in a complementary manner to satisfy the undertaken learning task [12]: instead of using manually designed or handcrafted modality-specific features, via DL we can automatically learn and extract an embedded representation for each modality. Furthermore, here we focus on supervised multimodal fusion that joins information from different modalities for a classification or regression task. The literature agrees that understanding when, which and how modalities should be fused is the main challenge in MDL [1,5].

To answer the first question (when), Baltруšaitis et al. divide multimodal fusion in model-agnostic and model-based methods [1]. The former are feature-based or decision-based, being also named as early and late fusion, respectively. Early fusion consists in the integration of the raw multimodal data into a single feature vector which is used as input for the DL model. At its simplest form, early fusion consists in concatenating multimodal features, which can bring to a redundancy of information. Oppositely, late fusion consists in using multiple models trained on separate modalities, whose decisions are joint via an aggregation function or a meta-learner, aiming to augment the performance of each unimodal model [13]. Because model-agnostic fusions can suppress intra- or inter-modality interactions, in the last years researchers have mainly focused on model-based methods that allow intermediate fusion, also referred to as joint fusion, to occur directly inside the models. DL models transform raw inputs into higher-level representations through the alternation of linear and non-linear operations. With a MDL model, we obtain these representations for each modality which will be fused into a hidden layer creating a joint multimodal representation. Thanks to the fact that with DL we can have an end-to-end training of all the multimodal representation components and the fusion component, the training phase essentially consists in learning hierarchical representations from raw data, which give rise to an optimized intermediate-level fusion representation [14]. This is why we can consider joint fusion an evolution of early fusion. Furthermore, inside a joint model, to understand when the unimodal architectures should be joined and to find the fusion structure, the majority of the available work follows a meticulous handcrafted approach. For instance, Neverova et al. adopted a single fusion layer from which they implement a gradual fusion strategy to automatically comprehend when to fuse the modalities [15].

Let us now focus on the second challenge, which aims to understand which modalities should be fused and with which models. With the increasing number of high-performing deep architectures, the selection of the one best suited for each modality is a hard process. The choice of which modality to fuse is usually based on intuition and on a trial and error approach that, usually, first fuses similar modalities and then tries to integrate disparate modalities [1]. Furthermore, prior investigations on the data, as well as feature selection techniques, should allow us to find redundant information along the modalities [16]. Shifting the attention to the choice of which neural networks should be fused, the analysis of the literature shows that the deep architectures are usually chosen by analyzing the single modalities [4,6], and no optimal multimodal architecture selection is overtaken.

Focusing on the third question, in the literature there are various techniques studying how to obtain an embedded representation that helps the unimodal models in the desired tasks. In joint fusion, the state-of-the-art adopts two main methodologies [17,18]. The first is a simple but effective way that concatenates the extracted vectors from
Take a given layer of each unimodal model \([17]\). The next hidden layer combines the different modalities by computing

\[
f(\sum_{i=1}^{m} w_i v_i)
\]

(1)

where \(v_i\) and \(w_i\) are the encoding and the weights of the model of the \(i\)th modality, \(f\) is the activation function and \(m\) stands for the number of modalities involved. The second approach consists of a multiplicative method computing the outer product \(\otimes\) (or Kronecker product) of all the \(m\) modality-specific feature vectors, extrapolated from a given hidden layer, exploiting the intra- and inter-modality relations [18]. The layer output is:

\[
\prod_{i=1}^{m} [v_i, 1]
\]

(2)

where the 1 is appended to the encoding \(v_i\) so that all unimodal and multimodal combinations interactions are represented. A schematic visualization of both joint fusion methods is shown in Fig. 1. Once the fusion occurs, the multimodal encodings are then fed to a fully connected (FC) classifier.

Many studies such as [19–21] dealt with determining an automatic methodology on how and when to fuse modalities, but none have focused on finding an optimized way for choosing which model architectures for each modality should be fused. Therefore, we hereby propose a strategy to find the best combination of models of the different modalities.

2.2. DL and COVID-19

The application of AI to the COVID-19 data analysis has been directed towards medical imaging and clinical data addressing two main issues: i) the detection of patients suffering from COVID-19 pneumonia from those which are healthy or affected by different types of pneumonia and, ii) the prediction of the outcome.

In the former case, a large number of publications have been produced, as reviewed in [22]. Many of them leverage the open-source CXR COVID-19 data collection [23], or specific clinical datasets. Moreover, there are also papers that use CT scans, as reviewed in [24], or that use other imaging modalities [22]. Instead, in the prediction scenario, few works have investigated the stratification of the disease to distinguish between patients with mild and severe outcomes. In this respect, the rest of this Subsection overviews such papers since outcome prediction is the focus of our work.

Considering the publications which work with CXR scans only, Signoroni et al. designed an end-to-end convolutional neural network (CNN) architecture for predicting a multi-regional score conveying the degree of lung compromised in COVID-19 patients [25].

Focusing on the publications using clinical data only, Zhu et al. presented a DL algorithm, whose architecture is composed of 6 FC dense layers with ReLU as activation function, which identifies the top clinical variable predictors and derived a risk stratification score system to help clinicians triage COVID-19 patients [26]. Their model was trained on a private dataset with 181 patients and of 56 clinical variables, obtaining an AUC equal to 0.968. Similarly, Al-Najjar and Al-Rousan built a neural network with one hidden layer to predict the status of recovered and death COVID-19 patients in South Korea [27]. They used seven different clinical variables and obtained an accuracy equal to 0.966 on a private dataset composed of 1308 patients.

As interpreting medical data is a multimodal process by its very nature, we now present the four works that use both clinical data and medical images, also noticing that three of them use CT images [28–30] and only one made use of CXR scans [4]. Fang et al. introduced a DL-based method using 53 clinical features and quantitative CT data to detect mild patients with a potential malignant progression [28]. The authors use a multi-layer perceptron (MLP) on the clinical features creating an embedding that is concatenated with the flattened CT, which are then fed to a Long short-term memory (LSTM) network and a FC network, achieving an accuracy of 0.792 on a cohort of 199 patients belonging to a private dataset. Next, Ning et al. used both CT images and 130 clinical features to discriminate between negative, mild and severe COVID-19 cases via a DL framework based on the VGG-16 and a 7-layer FC network [29]. This network was trained on a public dataset containing 1521 patients, and it obtained an accuracy equal to 0.811. Furthermore, Fang et al. developed an early-warning system with DL techniques to predict COVID-19 malignant progression leveraging CT scans and the clinical data of the patients [30]. The authors use a 3D ResNet and MLP to encode the chest CT, and then fed into an LSTM and several FC networks to make the prediction, obtaining an accuracy equal to 0.877, on a private dataset with 1040 patients.

Let us now turn the attention to the only work that, to the best of our knowledge, uses CXR images and clinical for the COVID-19 prognosis [4]. In this paper, the authors not only made publicly available a dataset of 1103 patients designed for the mild and severe stratification task, but they also presented three different multimodal AI approaches that can be used as a baseline performance for future improvement. In the first learning approach, also referred to as handcrafted approach (HC), the authors employed first order and texture features computed from the images, which are mined together with the clinical data feeding a supervised learner, such as Support Vector Machines and

![Diagram](image-url)
Random Forests. The second approach is named hybrid (HYB) and it mixes automatically extracted features computed by a pre-trained CNN with the clinical data, which are then fed to a classifier as before. The last approach is an end-to-end approach (ETE) exploiting together the clinical data and the raw CXR by using a multi-input CNN. The best results are obtained by the HYB approach, whose results are reported and discussed in Section 6.

This analysis of the literature shows that the development of AI-based models predicting the outcomes of COVID-19 patients still deserves further research efforts, since in the multimodal scenario only simple methods are employed with no interest in understanding when, which and how the neural architectures should be used to find the appropriate fusion. We decided to use [4] as our starting point, given the importance of their multimodal data, and we will show that our proposed MDL framework improves the performance available at the state-of-the-art.

2.3. XAI

The major disadvantage of neural network approaches is their lack of interpretability [31]. It is hard to understand what the predictions rely on, and which modalities and features hold an important role [32]. In general, an XAI algorithm is one that produces information to make a model’s functioning clear or easy to understand. This is why the field of XAI has grown in the last few years, with the development of many different methods suited for different tasks, as described in the overview offered by Holzinger et al. in [33]. The literature makes a clear distinction among models that are interpretable by design (transparent models), and those that can be explained by means of external XAI techniques (post-hoc explainability). Given the complex structure of DL algorithms, post-hoc model-specific XAI models have been developed. Focusing our attention to tabular or image data, i.e., the one used in this work, in the literature we found algorithms presenting feature relevance explanations specific for these modalities taken alone. For example, according to [31], Deep Taylor decomposition [34] and DeepLIFT [35] are the most used XAI models specific for multilayer neural networks which work on tabular data, Equalizer [36] and Grad-CAM [37] are suited for CNNs working on image data, whereas Integrated Gradients should be used by both network typologies [38].

The literature is well advanced in models which work on a single modality [31] but it lacks research for MDL. In this respect, the multimodal method that we present brings along the possibility to apply and combine any unimodal XAI algorithms in a way to understand the contribution that each modality and each data within a modality has during the prediction.

3. Materials

We performed the experiments on the ALforCOVID imaging archive [4] since it is the only publicly available multimodal dataset on COVID-19 stratification at the time this research was carried out. It includes 820 CXR scans and clinical data collected from six different hospitals at the time of hospitalization of symptomatic patients positive to the SARS-CoV-2 infection at the TR-PCR test. The 34 clinical available parameters are listed in Table A.1, and for further details the interested readers can refer to [4].

On the basis of the clinical outcome, each patient was assigned to the mild or severe group, where the mild group consists of the patients that were sent back to domiciliary isolation or were hospitalized without any ventilatory support, whilst the severe group is composed of patients who required non-invasive ventilation support, intensive care unit or patients who deceased.

Recently, a second release of the dataset was made available and it is composed of other 283 patients, whose data were collected in other two new centers. Here, we use these data as an external validation set in a way to have further proof that the proposed framework is robust on never seen data.

4. Methods

4.1. Multimodal learning

To predict the severity outcome of COVID-19 patients exploiting both the pixel information of their CXR scan and their clinical information, we propose a novel supervised end-to-end joint fusion method, which maximizes the advantages brought by all the given modalities. It first looks for the best combination of models, which are then joint and trained to carry out the given classification task. This is visually summarized in Fig. 2, which is organized in four main blocks that are now detailed, one per paragraph.

Unimodal learning. The proposed method can be applied to an arbitrary number of $m$ different modalities and $n$ different neural network architectures.

Let us first introduce the application matrix $\Theta \in \text{Mat}_{m \times n}$, where each element is defined as follows

$$
(\Theta)_{ij} = \begin{cases} 
1 & \text{if the } j\text{th model processes data of the } i\text{th modality} \\
0 & \text{if the } j\text{th model does not process data of the } i\text{th modality} 
\end{cases}
$$

Hence, $\theta_{ij}$ represents the use of the $j$th model for the $i$th modality, with $i \in \{1, \ldots, m\}$ and $j \in \{1, \ldots, n\}$. Furthermore, let us denote with $a_i$ any neural network architecture we would like to use; on this basis, when $\theta_{ij} = 1$, $a'_i$ denotes that the architecture $a_i$ is trained on data from the $i$th modality. The set of networks that will be combined is denoted by

$$
\Gamma = \{a'_i|\theta_{ij} = 1, \forall (i, j) \in I\}
$$

where $I$ contains a subset of all possible pairs of models and modalities:

$$
I \subseteq \{1, \ldots, m\} \times \{1, \ldots, n\}
$$

Consequently, the number of different configurations for $\Gamma_i$ is equal to $\Sigma_{s=1}^{m} \Sigma_{t=1}^{n} \theta_{st}$, where $s = \Sigma_{s=1}^{m} \Sigma_{t=1}^{n} \theta_{st}$. Furthermore, model architectures being part of $\Gamma_i$ depend on the corresponding $\Theta$ matrix, whilst the choice of the number of models $n$ and which models to include should consider the constraints on the applications, such as any limit in computational resources.

It is worth noting that the framework now introduced goes beyond the state-of-the-art of multimodal joint networks, which usually use only one model for each modality [6]. On the contrary, we consider the chance of having several models per modality, a situation that researchers and practitioners often meet in practice. In these cases, usually they first heuristically determine the best model per modality in a pool of many, i.e., $\Gamma_i$, and they combine such “best” models in a multimodal learner [31]. However, this approach neglects any possible interaction between models and modalities, so that the combination of best unimodal models does not necessarily reflect the best multimodal architecture. Furthermore, the general assumption that a multimodal model should be composed of a single model per modality is also a too much straightforward simplification.

Optimization. To go beyond the naive use of $\Gamma_i$, we deem that it should be more useful to work with all the modalities and all the models together. The first step is to detect which is the subset of models to be combined to get the best multimodal architecture providing the best multimodal data representation. This subset is determined by solving an optimization problem that leverages on two scores combined to get the best multimodal architecture providing the best multimodal data representation. This subset is determined by solving an optimization problem that leverages on two scores

$$
\ell(\Gamma_i)
$$

whilst the second is a measure of the diversity between the unimodal learning models, denoted as $D(\Gamma_i)$. To find optimal combination of models from $\Gamma_i$, exploiting at best all the modalities, denoted by $\Gamma^*$, the algorithm works as follows. Using a cross-validation approach, we first
apply each model $a_i'$, that was trained on the corresponding unimodal dataset $D_i$, to a validation set. Then, we compute the average of both $\mathcal{E}(T_j)$ and $D(T_j)$ among the cross-validation folds for all the possible $T_j$, finally finding the combination $G^*$ which maximizes both metrics. This means that we look for the $T_j$ that, on the one side, returns the best classification performance and, on the other side, reduces the incidence and effect of coincident errors among its members, thus considering possible relationships between models and modalities.

It is worth noting that such a process corresponds to solving a multi-objective optimization problem. In fact, given the pair $P = (F, C)$, where $F = \{f_1(x), \ldots, f_c(x)\}$ is the set of objective functions (a) that we want to maximize and $C = \{c_1(x) \leq 0, \ldots, c_b(x) \leq 0\}$ are the constraints (b), where if $x \in C$ it is said to be an admissible point for $P$. Given two admissible points $x_i$ and $x_j$ for $P$, $x_i \succ x_j$ means that $x_i$ dominates $x_j$. This means that $f_a(x_i) \leq f_a(x_j)$ $\forall a \in \{1, \ldots, a\}$, and $f_b(x_i) < f_b(x_j)$ for at least one $a \in \{1, \ldots, a\}$. A Pareto optimum $x^*$ is an admissible point for $P$ if $x^* \succ x$ for any other admissible points $x$ for $P$. i.e., it tries to minimize all functions in $F$ under the constraints $C$, solving a multi-objective minimization problem over $P$. Based on [42], the point $x^*$ is a Pareto optimum if there exist $\lambda$ and $\mu$ satisfying the following system:

$$
\begin{aligned}
\sum_{p=1}^{c} \lambda_p V_f(x^*) + \sum_{b=1}^{b} \mu_p V_c(x^*) &= 0 \\
\sum_{p=1}^{c} \lambda_p V_f(x^*) &= 0 \\
\lambda_p &\geq 0 \forall a = 1, \ldots, a \\
\mu_p &\geq 0 \forall b = 1, \ldots, b \\
(\lambda, \mu) &\neq (0, 0)
\end{aligned}
$$

In our framework, $G^*$ is the Pareto Optimum maximizing both two scores, related to the confusion matrix and to the diversity, respectively. This implies that $\mathcal{E}(T_j)$ and $D(T_j)$ are the two objective functions ($a = 2$). Thus, $G^*$ is a Pareto optimum if the quintuple $(\lambda_1', \lambda_2', \mu_1', \mu_2', G^*)$ satisfies the following system:

$$
\begin{aligned}
\lambda_1(-\mathcal{E}(T_j)) + \lambda_2(-D(T_j)) + \mu_1 - \mu_2 &= 0 \\
\lambda_1 \geq 0, \lambda_2 \geq 0, \mu_1 \geq 0, \mu_2 \geq 0
\end{aligned}
$$

where $\mathcal{E}(T_j)$ and $D(T_j)$ denotes the first derivative of such quantities. Observing that both $\mathcal{E}(T_j)$ and $D(T_j)$ range in $[0, 1]$, and under the hypothesis that $\mathcal{E}(T_j) = 1$ in case of perfect outcome prediction\(^1\) we can write the following minimization function:

$$
\min((\mathcal{E}(T_j) - 1)^2 + (D(T_j) - 1)^2)
$$

\(^1\) Straightforwardly, if $\mathcal{E}(T_j)$ is an error-related measure, i.e., $\mathcal{E}(T_j) = 0$ in case of perfect outcome prediction, Eq. (8) can be rewritten as $\mathcal{E}(T_j)^2 + (D(T_j) - 1)^2$.

Following the Kuhn–Tucker theorem [43], solving the minimization of Eq. (8) is equivalent to solving the system:

$$
\begin{aligned}
2(\mathcal{E}(T_j) - 1)\mathcal{E}(T_j) + 2(D(T_j) - 1)D(T_j) + \mu_1 + \mu_2 &= 0 \\
\mu_1 \geq 0, \mu_2 \geq 0
\end{aligned}
$$

(9)

$$(a_1', a_2', \mu_1', \mu_2', G^*)$$ satisfies the system and $G^*$ is a Pareto optimum for our optimization problem, when $a_1' = 2(1-\mathcal{E}(T_j))$ and $a_2' = 2(1-D(T_j))$, where $a_1' \geq 0$ and $a_2' \geq 0$ since $0 \leq \mathcal{E}(T_j) \leq 1$ and $0 \leq D(T_j) \leq 1$.

Interested readers in understating how to resolve the system in Eq. (9), can refer to [44].

In practice, with this optimization approach we want to find the combination of models, that when fused together maximize both an evaluation and a diversity metric, surpassing the constraint that if a modality is useless for the given classification task, all the networks working on that modality may be discarded.

Joint-late fusion. So far we presented an approach to find the optimal multimodal combination model, i.e., which modalities and related models should be included into the multimodal architecture, but we have not defined how and when to actually join the models $a_i'$ in $T_j$. To simplify the notation, we now omit the apex $i$, denoting the learners included in $T_j$ as $a_j$, with $j = 1, \ldots, |T_j|$. Each of these models provides a classification vector $y_j = [y_{j1}, \ldots, y_{jp}, \ldots, y_{jp}]^T$, where $c$ is the number of classes and $p$ is the generic $p$th output neuron. Each entry $y_{jp}$ is defined using the parameterized softmax function:

$$
y_{jp} = \text{softmax}(z_{jp}^{(0)}) = \frac{e^{z_{jp}^{(0)}}}{\sum_{i=1}^{c} e^{z_{ip}^{(0)}}}
$$

where $k > 0$, and $z_{jp}^{(0)}$ denotes the $p$th value of the output layer neuron internal activity for the network $a_j$; similar definition holds for $z_{jp}^{*}$. Note that for $k = 1$ each element represents an estimate of the classification posterior probability for each class of each model.

On these premises, to join the models we define a soft shared representation given by $[y_{j1}^T, \ldots, y_{j0}^T, \ldots, y_{jp}^T]^T$, which corresponds to concatenate the classification vectors, an approach similar to concatenating the hidden layers as in [17]. Hence, this shared representation vector is composed of $c \cdot |T_j|$ elements. Furthermore, if $k \to \infty$, $y_j$ becomes a binary vector with only one element equal to one, allowing us to get a crisp shared representation containing the binary outputs, which provides a direct interpretation of the embedded layer.

Classification layer. This shared representation, being either soft or crisp, is then passed to a FC neural network, with $c$ neurons in the last layer, that performs the intermediate fusion. Note that here we exploit the advantages of both joint and late fusion, since the classifications of
the sub-networks are aggregated via an end-to-end manner using the back-propagation during the training process that minimizes the overall loss function. Indeed, it avoids the combination of features belonging to different spaces, while it combines the deep neural networks into a common classification space, and it is shortly referred to as to joint-late fusion in the following.

4.2. XAI

The nature of the joint-late fusion approach, makes interpretability simple and effective. First, the weights $w$ coming out of the embedded classification vector connected to the first layer of the FC network, provides useful information because it reveals the importance of each class for each model and, consequently, the importance of each modality. The higher the value of $w_i$ with $i = 1, \ldots, n$, the higher is the contribution for the combined classification coming from $i$th model. Furthermore, as more models per modality are allowed, we can also understand the models that contribute more in the final classification for each class by looking at the weights proportion within each modality.

It is worth noting that any XAI algorithm, being model-agnostic or model-specific, can be applied to our MDL architecture, because it is composed of multiple models which can be interpreted one by one. For instance, if one of the modalities uses tabular data and the chosen models are FC neural networks, any post-hoc model-specific XAI algorithm for multi-layer neural networks can be applied on the single models to understand the importance of each tabular variable [34,35]. Moreover, if another modality consists of images and the models used are CNNs, we can use a post-hoc model-specific XAI algorithm to deduce the contribution of each image’s pixel to the classification [36,37]. Going forward, there are also XAI algorithms that work on models regardless of the used modalities, that when applied to CNNs or FC networks, we understand the contributions for the classification of each image pixel and of each tabular variable, respectively [38]. Since the model we propose is composed of multiple neural networks (both FC and CNNs as presented in Section 5), we can apply all the before-mentioned XAI algorithms on the corresponding network and sub-networks.

Given the fact in our framework we can have multiple models working with the same modality, we now introduce the concept of weighted XAI: it combines the outputs of the XAI algorithms specific to a modality with the weights coming from the embedded classification vector. Since these weights make us understand the contribution of each model, we combine the different XAI importances via a weighted sum to obtain a single interpretation of the modality. Hence for each feature $e_i$, the output of the weighted XAI algorithm importance is:

$$e_i = \sum_{n=1}^{\Psi} \omega_i e_i \omega_i$$

(11) where each contribution $e_i$ extracted from the $n$ unimodal models is weighted with the corresponding weight $\omega_i$ and summed together. For example, if we are working on the image modality and we apply Grad-CAM [37] to all the corresponding CNNs, we obtain a feature map for each model. By computing the dot product between the maps and the weight vector, we get a single feature map explaining which pixels contributed more or less to that specific classification. The same considerations hold for tabular data.

5. Experimental configuration

In this Section we introduce the different experiments we performed. Furthermore, to avoid any bias when comparing the baseline results of [4] that released the AIforCOVID dataset with the ones achieved by the learning framework presented here, we applied the same pre-processing procedure and validation approach, shortly described in Sections 5.1 and 5.5, respectively.

5.1. Data pre-processing

We use the 34 clinical descriptors available with the AIforCOVID repository, which are not direct indicators of the prognosis, that are listed in the appendix Table A.1. Missing data were imputed using the mean and the mode for continuous and categorical variables, respectively. Finally, to have the features all in the same range, a min–max scaler was applied along the variables.

In the case of CXR scans we extracted the segmentation mask of lungs, using a pre-trained U-Net [45] on two non-COVID-19 datasets [46,47]. The mask was used to extrapolate the minimum squared bounding box containing both lungs, which is then resized to $224 \times 224$ and normalized with a min–max scaler bringing the pixel values between 0 and 1. As already mentioned, the interested readers can refer to [4] for further details.

5.2. Classification

As introduced in Section 4, the first step to follow to obtain the optimal $\Gamma^*$ network is to train and evaluate different models using the data of the two modalities, i.e., CXR scans and clinical data. For the image modality we worked with 30 different CNNs that come from 8 different main architectures:

- **AlexNet** [48];
- **VGG** [49]: VGG11, VGG11-BN, VGG13, VGG13-BN, VGG16, VGG16-BN, VGG19, VGG19-BN;
- **ResNet** [50]: ResNet18, ResNet34, ResNet50, ResNet101, ResNet152, ResNetXt50, ResNetXt101, Wide-ResNet50-2, Wide-ResNet101-2;
- **DenseNet** [51]: DenseNet121, DenseNet169, DenseNet161, DenseNet201;
- **GoogLeNet** [52];
- **ShuffleNet** [53]: ShuffleNet-v2-x0-5, ShuffleNet-v2-x1-0, ShuffleNet-v2-x1-5, ShuffleNet-v2-x2-0;
- **MobileNetV2** [54];
- **MNASNet** [55]: MNASNet0-5, MNASNet1-0;

In all the cases the weights were initialized using the values pretrained on the ImageNet dataset [56]; we also changed the output layer dimension to 2 neurons, one for each class.

In the case of clinical information, which are tabular data, we adopted 4 MLPs that differ in terms of depth and wideness of the model. We opted to use such architectures since these feedforward networks are able to learn a low-dimensional representation before being fused with the other modality [57]. In particular, the models’ hidden layers have the following organizations:

- **MLP-1**: it has 3 hidden layers with 64, 64, 32 neurons respectively;
- **MLP-2**: it has 5 hidden layers with 64, 128, 128, 64, 32 neurons respectively;
- **MLP-3**: it has 7 hidden layers with 64, 128, 256, 256, 128, 64, 32 neurons respectively;
- **MLP-4**: it has 9 hidden layers with 64, 128, 256, 512, 512, 256, 128, 64, 32 neurons respectively;

A ReLU activation function is applied on all layers, since it learns several times faster than regular sigmoid activation functions [58], despite its non-differentiability at zero. Straightforwardly, the input layer and the output layers consist of 34 and 2 neurons, respectively. Turning our attention to the matrix $\Theta$, it is composed of 24 columns ($n = 24$ the number of architectures) and two rows ($m = 2$ the number of modalities).

Furthermore, as for the two objective functions $\mathcal{L}(\Gamma)$ and $\mathcal{D}(\Gamma)$ we adopt the accuracy (Acc) and the correlation coefficient ($\rho$), respectively. Note that the accuracy is computed on the models part of the $\mathcal{I}$’s in late-fusion via the majority voting aggregation function. We opted for late fusion optimization to prevent the training of all the $\sum_{\mathcal{I},\mathcal{J}}$ end-to-end ensemble models. In this way, we only train $n$ single models, apply the optimization, and conclude with one final embedded training of $\Gamma^*$. 
Majority voting consist in finding the most common classification $c_i$ extracted by the $|I|_i$ unimodal trained models

$$c = M(o(c_i), i = 1, ... , |I|) $$

where $|I|$ indicates the mode of a set and $|I|$ is the number of models in $I_i$. The correlation coefficient $\rho$ instead between two classifiers is defined as:

$$\rho_{ij} = \frac{n^{11}n^{00} - n^{01}n^{10}}{\sqrt{n^{11}n^{01} + n^{01}n^{10} + n^{10}n^{00}}}$$

where $n^{11}$ is the number of correct classifications made by both models, $n^{10}$ is the number of correct classifications made by the models, $n^{10}$ and $n^{0}$ are the number of instances on which the classifications of the two models differ. Since $\rho$ is a pairwise measure, the overall diversity in $I_i$ is given by:

$$\rho(I_i) = \frac{2}{|I_i||I_i| - 1} \sum_{j \in I_i} \sum_{j \in I_i} \rho_{ij}$$

We opted on using these evaluation and diversity metrics, given that they are metrics that focus on classified observations both for positive and negative, and on correctly and incorrectly instances, respectively.

Going beyond the single models, after determining $\Gamma^*$, its models are joint via the joint-late fusion method introduced in . As mentioned there, the classification space can be built adopting a soft and a crisp shared representation. This permits us to investigate two learners, denoted as JLF-S and JLF-C, respectively. We also investigate what happens by varying the final FC layer, lying after the shared representation. To this end, we tested two different FC layers: the first does not have any hidden layers but directly connects the joint vector to the output classification layer using a single perceptron, and it is denoted using the ending 1. The second consists of a single hidden layer of 4 neurons connected to the output classification layer, which is denoted using the ending 2. The former permits us to investigate linear combinations among the classifications of the networks, whilst the latter investigates non-linear combinations. In summary, we investigate four set-ups for the joint-late approach we present, i.e., JLF-S-1, JLF-S-2, JLF-C-1 and JLF-C-2.

5.3. Training configuration

For both the single model and the multimodal scenarios, we adopt the training procedure described in [4]; so that any variation in performance is given only by the enhancements brought by the method and not by any other reason. To prevent overfitting of the CNNs, we randomly applied the following image transformations with a probability equal to 0.5: horizontal or vertical shift ($\leq 20$ pixels $\leq 20$), random zoom ($0.9 \leq \text{factor} \leq 1.1$), vertical flip, random rotation ($-15^\circ \leq \text{angle} \leq 15^\circ$), and elastic transform ($20 \leq a \leq 40$; $\sigma = 7$). Both the CNNs and the MLPs were trained using the cross-entropy loss, regulated by an Adam optimizer with an initial learning rate of 0.001, which is scheduled to reduce by an order of magnitude every time the minimum validation loss does not change for 10 consecutive epochs. To prevent overtraining and overfitting we fix the number of maximum epochs to 300, with an early stopping of 25 epochs on the validation loss. We do not perform any preliminary optimization of the networks hyperparameters because in Arcuri and Fraser empirically reported that fine-tuning the learner parameters provides performance that do not statistically differ from those attained using the default values [59], a claim also confirmed by the experiments presented in [4] on the AIforCOVID dataset. Furthermore, we do not pre-train the networks on another CXR dataset: indeed, although this would help the models learn modality-specific feature representations [60], the results available on the AIforCOVID dataset show that this practice does not introduce any significant performance improvement [4].

All the experiments were trained by using a batch size of 32 on a NVIDIA TESLA V100 GPU with 16 GB of memory, using PyTorch as the main DL coding library.

5.4. Competitors

To verify that the end-to-end process actually brings a performance enhancement, and to check that the joint network is not actually learning a late-fusion via FC layers using as input the classifications of the different networks, we froze the joint networks up to the joint classification vector and trained only the last FC classifier, creating a late fusion method. We would expect that if there is an actual benefit in the joint end-to-end to fusion, these performances would be lower. We will denote these experiments with: LF-S-1, LF-S-2, LF-C-1 and LF-C-2, where the S and the C encodings refer to the soft and the crisp embedded classification vectors and 1 and 2 refer to the type of FC layers employed, as before.

We also compare the results of the joint-late fusion with those attained by the majority voting late fusion approach applied to the models included in $\Gamma^*$, denoted as LF-MV. Since majority voting is the aggregation function used in the optimization phase, this comparison permits us to not only investigate if the joint-late training is beneficial, but also to compare our proposal against a late-fusion technique, addressing the question on when the models should be fused. Therefore, we also compare our method with an early fusion method, denoted as EF, using a Support Vector Machine for the classification, which receives as input the concatenation of the extracted features from the last hidden layers of the networks in $\Gamma^*$. We also tested the concatenation (denoted as JF-C) and the multiplication methods (denoted as JF-M) introduced in Section 2.1: the comparison of these approaches against the JLF solutions lets us deepen how the single models should be aggregated and at which level the fusion should occur. For the sake of completeness, we would note that the JLF approach intrinsically studies which learners should be fused and, in the next Section 6 we also compare the results against the use of a single network.

To validate the usefulness of both the evaluation and diversity metric to find $\Gamma^*$, we adjust the minimization function of Eq. (8) by considering only one of the metrics at a time. We take the best performing JLF model and obtain JLF$^*$ and JLF$^{**}$, which denote $\Gamma^*$-optimized only on $\epsilon$ and $D$, respectively.

Finally, as natural competitors, we also compare our results with those presented in the work introducing the AIforCOVID dataset [4], i.e., the HC, HYB and ETE methods summarized in Section 2.

Note that for all the experiments where joint training was performed (i.e., JLF-C-1, JLF-C-2, JLF-S-1, JLF-S-2, JF-C, JF-M, JLF$^*$, JLF$^{**}$), the weight initialization is performed by using the weights of the networks trained singularly.

5.5. Validation approach

The experiments were ran in 10-fold stratified cross-validation (CV), and leave-one-center-out cross-validation (LOCO), following the same experimental procedure described in [4], thus ensuring a fair competitor between the approaches. In each CV fold the proportion between the training, validation and testing sets is 70%-20%-10%, respectively.

In LOCO validation we study how the models generalize to different data sources: indeed, in each fold the test set contains all the samples belonging to one center only, while the instances of the other six centers were assigned to the training and validation set.

With reference to [4], here we add another layer of external validation (EV), ran on an external dataset. To this goal, we exploited the second release of the AIforCOVID dataset, which contains data collected from 283 patients belonging to other two centers, as already mentioned in Section 3.

5.6. Performance metrics and statistical assessments

The accuracy, the sensitivity and the specificity are the evaluation metrics used to assess the performance, as in [4].
is interpreting the data in the correct manner via the XAI methods all give useful and distinct information for the prognosis task, since (Acc), sensitivity (TPR) and specificity (TNR), with the respective statistics shown in Table 1, which shows the average accuracy for each modality to a maximum of 3.

Furthermore, we apply the one-way ANOVA among the different groups of models and, to interpret the statistical significance, we used the pairwise Tukey test with a Bonferroni p-value correction at $a = 0.05$. Henceforth, we refer to a statistically significant difference when this test is satisfied.

6. Results and discussions

In this Section we first show the results of the aforementioned experiments for the classification task, then we verify if the model is interpreting the data in the correct manner via the XAI methods introduced in Section 4.

6.1. Classification task

Before presenting the numerical results, let us report that the Pareto optimum $I^*$ is composed of three CNNs and one MLP, namely the GoogLeNet, the VGG13-BN, the ResNet50-32 × 4d, and the MLP-2. It is worth noting that the selected models belong to different families suggesting that each extrapolates different information to satisfy the desired classification. Moreover, we understand that the two modalities all give useful and distinct information for the prognosis task, since we have at least one model for each modality. Furthermore, we also ran the multi-objective optimization fusing the learners in each $I_\ell$ applying end-to-end training rather than late fusion to investigate if any differences exist. 2 The results reveal that both provided the same optimal $I^*$, confirming that performing the optimization in late fusion rather than in end-to-end fashion is a viable alternative that strongly reduces the computation time since the former does not require any additional training.

The quantitative results achieved by all the learners presented in Section 5 are shown in Table 1, which shows the average accuracy (Acc), sensitivity (TPR) and specificity (TNR), with the respective standard deviations, for the CV, LOCO, and EV results. The

| Proposal | Model | CV Acc | TPR | TNR | LOCO Acc | TPR | TNR | EV Acc | TPR | TNR |
|----------|-------|--------|-----|-----|----------|-----|-----|--------|-----|-----|
| JLF-C-1  | 79.75 ± 0.23 | 82.47 ± 0.37 | 76.81 ± 0.24 | 77.86 ± 0.33 | 79.95 ± 0.17 | 75.95 ± 0.30 | 77.61 ± 1.10 | 79.18 ± 1.21 | 75.38 ± 1.11 |
| JLF-C-2  | 79.63 ± 0.24 | 80.99 ± 0.17 | 78.90 ± 0.23 | 77.30 ± 0.32 | 79.06 ± 0.22 | 75.88 ± 0.17 | 78.02 ± 1.22 | 79.50 ± 1.26 | 74.23 ± 1.23 |
| JLF-S-1  | 79.63 ± 0.16 | 82.47 ± 0.15 | 76.67 ± 0.20 | 77.39 ± 0.29 | 80.00 ± 0.27 | 75.52 ± 0.14 | 77.90 ± 1.27 | 78.79 ± 1.24 | 75.96 ± 1.21 |
| JLF-S-2  | 79.39 ± 0.17 | 81.43 ± 0.16 | 76.69 ± 0.25 | 77.16 ± 0.30 | 79.98 ± 0.25 | 75.89 ± 0.31 | 76.54 ± 1.25 | 79.07 ± 1.23 | 74.73 ± 1.16 |
| LF-MV   | 78.41 ± 0.32 | 80.09 ± 0.31 | 76.24 ± 0.22 | 75.48 ± 0.29 | 77.98 ± 0.20 | 73.82 ± 0.21 | 74.88 ± 1.13 | 77.64 ± 1.14 | 75.31 ± 1.15 |

| Proposal | Model | CV Acc | TPR | TNR | LOCO Acc | TPR | TNR | EV Acc | TPR | TNR |
|----------|-------|--------|-----|-----|----------|-----|-----|--------|-----|-----|
| JF-C     | 72.80 ± 0.63 | 76.62 ± 0.83 | 68.62 ± 0.98 | 70.35 ± 0.47 | 71.04 ± 0.52 | 67.95 ± 0.73 | 72.47 ± 1.29 | 78.76 ± 1.44 | 68.24 ± 1.39 |
| JF-M     | 75.00 ± 0.76 | 78.30 ± 0.71 | 71.47 ± 0.85 | 71.37 ± 0.49 | 74.59 ± 0.65 | 70.47 ± 0.49 | 71.64 ± 1.37 | 75.06 ± 1.41 | 69.90 ± 1.83 |
| JLF$^a$  | 76.34 ± 0.13 | 78.39 ± 0.18 | 73.62 ± 0.23 | 74.16 ± 0.26 | 76.91 ± 0.29 | 72.76 ± 0.12 | 73.51 ± 1.29 | 79.63 ± 1.26 | 71.42 ± 1.12 |
| JLF$^b$  | 75.31 ± 0.16 | 77.41 ± 0.19 | 72.67 ± 0.25 | 73.20 ± 0.30 | 75.89 ± 0.30 | 71.88 ± 0.09 | 72.58 ± 1.22 | 79.07 ± 1.27 | 70.74 ± 1.15 |
| EF       | 75.94 ± 0.45 | 77.89 ± 0.36 | 73.67 ± 0.93 | 73.56 ± 0.29 | 75.78 ± 0.90 | 67.77 ± 0.76 | 71.90 ± 1.52 | 73.46 ± 1.16 | 65.87 ± 1.47 |
| HC       | 75.50 ± 0.70 | 75.80 ± 0.80 | 75.30 ± 1.30 | 75.20 ± 0.70 | 71.10 ± 1.60 | 82.40 ± 1.50 | 75.34 ± 1.76 | 69.21 ± 17.52 | 80.32 ± 16.65 |
| HYB      | 76.90 ± 5.40 | 78.80 ± 6.40 | 74.70 ± 5.90 | 74.30 ± 6.10 | 76.90 ± 18.90 | 68.50 ± 15.50 | 72.76 ± 7.53 | 75.43 ± 18.24 | 66.43 ± 16.42 |
| ETE      | 74.80 ± 0.80 | 75.40 ± 1.70 | 75.10 ± 1.50 | 70.90 ± 0.50 | 73.40 ± 1.80 | 69.60 ± 0.90 | 68.89 ± 1.53 | 71.53 ± 1.52 | 67.65 ± 1.60 |

2 For computational restrictions during end-to-end training, we limited the number of possible architectures in $I_\ell$ for each modality to a maximum of 3.

reveal that the four joint-late methods, i.e., the first four rows in Table 1, outperform the other techniques, and the performance differences against the other multimodal approaches are statistically significant in terms of accuracy, sensitivity and specificity. In particular, these significant differences between JLF and LF-MV, LF-P and LF-D methods demonstrate that the end-to-end training of the method is effective and provide better results than the late fusion of the networks in the ensemble. As mentioned in Section 5.4, this result displays when the models in the ensemble should be fused.

Let us now focus on the comparison between the four JLF methods and the other two approaches providing an embedded multimodal representation, namely the JF-C, JF-M competitors described in Section 2. We find that the JLF performance are statically different in terms of accuracy, sensitivity and specificity with respect to those attained by the JF-C, JF-M techniques. This finding points out how the individual learners should be combined to provide a useful embedded multimodal representation.

We also found that the performance of JLF models are statistically different when compared with those presented in [4] (i.e., HC, HYB and ETE), which are the current baseline for the AlforCOVID dataset.

Turning our attention to the comparison between the four set-ups of the JLF approach, the results reveal that no statistically significant differences exist among their performance, suggesting that there is no preferred technique between them, while JLF-C1 provides the best accuracy in both CV and LOCO experiments.

We are now interested in assessing to what extent our optimization algorithm effectively detects $I^*$ as the best choice of which learners should be combined among all the possible $I_\ell$. To this end, we randomly picked 100 different $I_\ell$ that are joint using the JFL-C1 technique for the reason mentioned a few lines above. Next, we compare these performance against the four JLF results in Table 1 that are achieved using $I^*$, and we find that they are statistically significantly lower.

We now focus on the unimodal scenario to investigate the performance in JLF when a modality is missing: to this end Table 2 shows the results of single models in $I^*$ when trained on a single modality.
By performing the statistical assessments we find that there is no significant change with the unimodal results presented in [4], as it could be expected. Interested readers can refer to Table B.2 which reports all the results attained by all the unimodal CNNs and MLPs considered. Nevertheless, GoogLeNet, VGG13-BN and ResNeXt50 are actually the first, the second and the sixth, among the CNNs, respectively, whilst MLP-2 is the first among the other MLPs. Comparing the four JLF learners with these unimodal networks, we find that the former provide performance that are always statistically larger than those returned by the latter, confirming the importance of the fusion between the modalities.

Let us now concentrate on a single modality. Since \( \Gamma^* \) is composed of three CNNs working on the image modality, we perform an ablation test removing the network using clinical data (i.e., MLP-2). In this case, JLF-C-1 achieves an average accuracy equal to 76.09 ± 0.29, 75.68 ± 0.31 and 74.86 ± 0.27 in CV, LOCO and EV, respectively. Since these values are statistically lower than those of JLF, we deem that joining multiple networks is beneficial. Furthermore, note that the apposite ablation test, i.e., removing the networks using image data, corresponds to consider MLP-2 only, whose performance are already shown in Table 2.

As a further observation, let us now focus on the different performance achieved in 10-fold CV, LOCO and EV. To make the comparison easier, Fig. 3 shows the distribution of the performance of the different methods across all the experimental scenarios. Observing the results of the JLF methods, we notice that the method is robust and generalizable since the performance drop in LOCO and in EV is limited, while all the others suffer from a larger decrease.

The main limitation of this method is its computational cost, given that all the unimodal models and the resulting joint model need to be trained. As the number of models and the number of modalities increases, the complexity rises. One may argue that it may be more practical to spend time and resources to optimize a single model per modality, but Table 2 shows that there are unimodal networks which perform poorly, which would bring to a sub-optimal solution. The goal of our method is not to find the best performing single model but to find the optimal fusion from a pool of classifiers, making it very useful in real-world scenarios where it is hard to know a priori which models are best suited for a specific task. Furthermore, our method agrees with the no free lunch theorem [61], which states that the computational cost of finding a solution, averaged over all problems in the class, is the same for any solution method.

**Explainability.** To open the black-box nature of the multimodal deep approach presented in this work, we apply the XAI algorithms introduced in Section 4 to the model with the highest performance, i.e., JLF-C-1.

To this end, we first extract the weights coming out of the classification vector which are connected to the next hidden layer of the FC network; second, we compute the mean relative intensities of such weights attained in CV, obtaining the following distribution: 17%, 26%, 16%, and 41% for GoogLeNet, VGG13-BN, ResNeXt50, and MLP-2, respectively. This information not only makes us understand the importance of every single model for the final classification, but it also explains the hierarchy inter- and intra-modality. In particular, we notice that the image modality has more importance than the clinical one since the relative vector weights are 59% and 41%, respectively. Even if there is a discrepancy between the number of models per modality, it is interesting to notice that MLP-2 has the highest importance among the single neural networks, but the combination of the three CNNs gives higher relative importance to the image modality. Going deeper, we can see that there is also an intra-modality hierarchy between the image-based models, where by calculating the relative weight importances on GoogLeNet, VGG13-BN, and ResNeXt50, we obtain 29%, 44%, and 27%, respectively.

To enable physicians to explore and understand data-driven DL-based systems, we work on XAI algorithms for single models. For each model composing the joint fusion, we apply XAI algorithms suited for the specific modality. For example, Fig. 4 shows the results of the integrated gradients algorithm [38] applied to MLP-2, which shows the importance of the clinical features for the classification. Observing the results on the test sets, we notice that the most important features are the difficulty in breathing and the oxygen percentage in blood. Indeed, the presence of difficulty in breathing and lower values in the oxygen percentage in the blood are indications of the severe COVID-19 cases as also the medical literature confirms [62]. Going forward, we can use the aforementioned relative weights of the classification vector for a specific modality to combine the results coming from an XAI algorithm. Still in Fig. 4, we show the feature maps extracted from GoogLeNet, VGG13-BN, ResNeXt50 by applying Grad-CAM [37], and by combining them via Eq. (11) we obtain a resulting map for the modality. In this way we understand how the different components working on the image modality, and as a whole, interpret the pixel values. Many
more XAI algorithms can be applied following this methodology, but to prevent redundancy we limit ourselves to these examples. We did not perform any evaluation on the XAI methods, since it is out of our scope of the presented method.

7. Conclusions

In this manuscript we have presented an approach to build an optimized multimodal end-to-end model exploiting a performance metric and the diversity of different neural networks. It addresses open issues in MDL related to when, which and how the modalities and the related learners must be joined to provide effective embedded representation and satisfactory performance. Given the impact of the COVID-19 pandemic, we applied our method to predict patients’ prognosis, a topic that has attracted recent research after the large efforts towards the detection of COVID-19 signs in medical images [22]. In this scenario, the use of different modalities capturing different aspects of the disease progression should be useful and, for this reason, we used the AlforCOVID dataset [4], which is the only publicly available dataset containing clinical data and CXR scans. Our proposal algorithmically determines which deep architectures for each modality should be fused, by exploiting Pareto Optimization. Furthermore, here we have investigated how and when the fusion of the modalities should occur: we find that concatenating the classification vectors of each learner, being either soft or crisp labels, is a viable solution. In terms of performance, our method attains state-of-the-art results, not only outperforming the baseline performance reported in [4] but also being robust to external validation. Moreover, our method fits well with many XAI algorithms, which allow us to figure out a hierarchy among modalities and they can extract the intra-modality importance of the different features by combining the results of various networks.

Future directions are directed towards five main goals. First, we plan to study a method to minimize the computational costs to extend the number of possible architectures per modality in the case of end-to-end training. Second, since the presented method is not specific to medical data, a further investigation to other multimodal datasets of other fields is needed. Third, to demonstrate the generalizability of the proposed method, future work is directed also on considering datasets composed of more modalities. It is worth noting that there exists several optimization frameworks proven to be very practical healthcare, such as monarch butterfly optimization [63], earthworm optimization algorithm [64], elephant herding optimization [65], moth search [66], slime mould algorithm [67], hunger games search [68], Runge Kutta optimizer [69], colony predation algorithm [70], and Harris hawks optimization [71] and those based on data [72,73] and graph representations [74,75]. On this basis, our fourth direction of future investigation will investigate and compare different optimization methodologies. Lastly, as medical data should come from different institutions, it would also be interesting to explore the deployment of our approach in a federated learning framework.

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.

Acknowledgments

This work is partially funded by: i) POR CAMPANIA FESR 2014 - 2020, API-OS1.3 project “Protocolli TC del torace a bassissima dose e tecniche di intelligenza artificiale per la diagnosi precoce e quantificazione della malattia da COVID-19” CUP D54I20001410002; ii) EU project “University-Industrial Educational Centre in Advanced Biomedical and Medical Informatics (CeBMI) No. 612462-EPP-1-2019-1-SK-EPPKA2-KA”, iii) FONDO PER LA CRESCITA SOSTENIBILE (F.C.S.) Bando Accordo Innovazione DM 24/5/2017 (Ministero delle Imprese e del Made in Italy), under the project entitled “Piattaforma per la Medicina di Precisione. Intelligenza Artificiale e Diagnostica Clinica Integrata” (CUP B89J23000580005); iv) “PE013 - Future Artificial Intelligence Research (FAIR)”, supported by European Union — NextGenerationEU, PNRR Topic M4C2: Dalla ricerca all’impresa - I.3:Partenariati estesi a università, centri di ricerca, imprese e finanziamento progetti di ricerca di base, CUP C53C22008000006 (Univ. Campus Bio-Medico di Roma), CUP Master B53C22003630006. The views and opinions expressed are however only those of the authors and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them. The authors would like to thank the team collecting and making publicly available the AlforCOVID dataset.

Appendix A. Clinical data

See Table A.1.
### Table A.1
Description of the clinical data coming from [4]. Feature names followed by + were not used.

| Name                        | Description                                                                 |
|-----------------------------|-----------------------------------------------------------------------------|
| Active cancer in the last 5 years | Patient had active cancer in the last 5 years                              |
| Age                         | Patient’s age (years)                                                       |
| Atrial Fibrillation         | Patient had atrial fibrillation                                              |
| Body temperature (°C)       | Patients temperature at admission (°C)                                      |
| Cardiovascular Disease      | Patient had cardiovascular diseases                                          |
| Chronic Kidney disease      | Patient had chronic kidney disease                                          |
| COPD                        | Chronic obstructive pulmonary disease                                       |
| Cough                       | Cough presence                                                              |
| CRP                         | C-reactive protein concentration (mg/dL)                                    |
| Days Fever                  | Days of fever up to admission (days)                                        |
| D-dimer                     | D-dimer amount in blood                                                     |
| Death+                      | Death of patient occurred during hospitalization for any cause             |
| Dementia                    | Patient had dementia                                                        |
| Diabetes                    | Patient had diabetes                                                        |
| Dyspnea                     | Patient had intense tightening in the chest, air hunger, difficulty breathing, breathlessness or a feeling of suffocation |
| Fibrinogen                  | Fibrinogen concentration in blood (mg/dL)                                   |
| Glucose                     | Glucose concentration in blood (mg/dL)                                      |
| Heart Failure               | Patient had heart failure                                                   |
| Hypertension                | Patient had high blood pressure                                            |
| INR                         | International Normalized Ratio                                              |
| Ischemic Heart Disease      | Patient had ischemic heart disease                                          |
| LDH                         | Lactate dehydrogenase concentration in blood (U/L)                         |
| O₂ (%)                      | Oxygen percentage in blood (%)                                              |
| Obesity                     | Patient had obesity                                                         |
| PaCO₂                       | Partial pressure of carbon dioxide in arterial blood (mmHg)                 |
| PaO₂                        | Partial pressure of oxygen in arterial blood (mmHg)                        |
| PCT                         | Platelet count (mg/mL)                                                     |
| pH                          | Blood pH                                                                    |
| Position+                   | Patient position during chest x-ray (%supine)                              |
| Positivity at admission     | Positivity to the SARS-CoV-2 swab at the admission time                    |
| Prognosis                   | Patient outcome                                                             |
| RBC                         | Red blood cells count (10^9/L)                                              |
| Respiratory Failure         | Patient had respiratory failure                                             |
| SaO₂                        | Arterial oxygen saturation (%)                                              |
| Sex                         | Patient’s sex                                                               |
| Stroke                      | Patient had stroke                                                          |
| Therapy Anakinra+           | Patient was treated with Anakinra                                           |
| Therapy anti-inflammatory+  | Patient was treated with anti-inflammatory drugs therapy                    |
| Therapy antiviral+          | Patient was treated with antiviral drugs                                    |
| Therapy Eparine+            | Patient was treated with eparine treatment; therapeutic treatment          |
| Therapy hydroxychloroquine+ | Patient was treated with hydroxychloroquine                                 |
| Therapy Tocilizumab+        | Patient was treated with Tocilizumab                                        |
| WBC                         | White blood cells count (10^9/L)                                            |

Appendix B. Single models

See Table B.2

### Table B.2
Performance (Acc, TPR, TNR) of all the single modality models. The models with the highest accuracy are highlighted for each data modality (CLI, IMG) and validation scenario (CV, LOOCV, EC).

| Model   | Data   | CV  | LOOCV | EV  |
|---------|--------|-----|-------|-----|
|         |        | Acc |       |     |
|         |        | TPR |       |     |
|         |        | TNR |       |     |
|         |        |     | Acc   |     |
|         |        |     | TPR   |     |
|         |        |     | TNR   |     |
| MLP-1   | CLI    | 74.56 ± 0.32 | 71.42 ± 0.70 | 73.81 ± 0.22 |
| MLP-2   | CLI    | 74.87 ± 0.46 | 72.50 ± 0.16 | 73.91 ± 0.29 |
| MLP-3   | CLI    | 74.95 ± 0.53 | 72.66 ± 0.13 | 73.98 ± 0.34 |
| MLP-4   | CLI    | 73.93 ± 0.42 | 74.87 ± 0.36 | 72.77 ± 0.14 |
| AlexNet | IMG    | 55.97 ± 0.21 | 92.00 ± 0.00 | 58.51 ± 0.87 |
| VGG11   | IMG    | 58.17 ± 0.69 | 82.07 ± 0.71 | 58.51 ± 0.87 |
| VGG11-BN| IMG    | 70.24 ± 0.18 | 63.45 ± 0.11 | 67.16 ± 0.64 |
| VGG13   | IMG    | 61.70 ± 0.30 | 55.92 ± 0.42 | 54.56 ± 0.11 |
| VGG13-BN| IMG    | 73.04 ± 0.38 | 75.67 ± 0.23 | 70.86 ± 0.47 |
| VGG16   | IMG    | 58.65 ± 0.40 | 90.76 ± 0.81 | 58.97 ± 0.10 |
| VGG16-BN| IMG    | 70.85 ± 0.18 | 63.89 ± 0.99 | 77.59 ± 0.82 |
| VGG19   | IMG    | 55.73 ± 0.90 | 92.00 ± 0.00 | 58.51 ± 0.87 |
| VGG19-BN| IMG    | 68.78 ± 0.01 | 61.46 ± 0.23 | 67.14 ± 0.22 |
| ResNet18| IMG    | 72.56 ± 0.53 | 69.03 ± 0.31 | 75.59 ± 0.40 |
| ResNet34| IMG    | 72.92 ± 0.56 | 69.23 ± 0.37 | 76.06 ± 0.95 |

(continued on next page)
### References

1. T. Baltrusaitis, C. Ahuja, L.-P. Morency, Multimodal machine learning: A survey and taxonomy, IEEE Trans. Pattern Anal. Mach. Intell. 41 (2) (2018) 423–443.

2. J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, A.Y. Ng, Multimodal deep learning, in: Proceedings of the International Conference on Machine Learning, 2011, pp. 689–696.

3. D. Ramachandran, G.W. Taylor, Deep multimodal learning: A survey on recent advances and trends, IEEE Signal Process. Mag. 34 (6) (2017) 96–108.

4. P. Soda, N.C. D'Amico, J. Tessadri, G. Valbusa, V. Guarrasi, C. Cortobello, M.U.C. Abr, S. Ricci, E. Cordelli, D. Fazzini, et al., AlterCOVID: predicting the clinical outcomes in patients with COVID-19 applying AI to chest X-rays: an italian multicentre study, Med. Image Anal. 74 (2021) 102216.

5. S.R. Stahlschmidt, B. Ullenborg, J. Syngenring, Multimodal deep learning for biomedical data fusion: a review, Brief. Bioinform. 23 (2) (2022) bbb669.

6. R.J. Chen, M.Y. Lu, J. Wang, D.F. Williamson, S.J. Rodig, N.J. Lindeman, F. Mahmood, Pathomic fusion: an integrated framework for fusing histopathology, and genomic features for cancer diagnosis and prognosis, IEEE Trans. Med. Imaging 41 (4) (2020) 757–770.

7. L.I. Kuncheva, C.J. Whitaker, Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy, Mach. Learn. 51 (2) (2003) 181–207.

8. A. Holzingr, The next frontier: AI we can really trust, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2021, pp. 427–440.

9. V. Guarrasi, L. Tronchin, D. Albano, E. Faitella, D. Fazzini, D. Santucci, P. Soda, Multimodal explainability via latent shift applied to COVID-19 stratification, 2020 IEEE preprint arXiv:2002.12174.

10. V. Chamola, V. Hassija, V. Gupta, M. Guizani, A comprehensive review of the COVID-19 pandemic and the role of IoT, drones, AI, blockchain, and 5G in managing its impact, IEEE Access 8 (2020) 90225–90265.

11. C.M. Garau, V. Guarrasi, E. Cordelli, R. Sicilia, S. Gentile, L. Messina, M. Fiore, C. Piccolo, B. Beomonte Zobel, G. Iannello, et al., A multimodal ensemble driven by multiobjective optimisation to predict overall survival in non-small-cell lung cancer, J. Imag. 8 (11) (2022) 298.

12. A. Holzingr, M. Dehmer, F. Emmert-Streib, I. Augenstein, J. Del Ser, W. Samek, I. Jurisica, N. Díaz-Rodríguez, Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artificial intelligence, Inf. Fusion 79 (2022) 265–278.

13. Y. Wu, J.J. Arribas, Fusing output information in neural networks: Ensemble performs better, in: Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Vol. 3, IEEE, 2003, pp. 1692–1706.

14. Y. Bengio, A. Courville, P. Vincent, Representation learning: A review and new perspectives, IEEE Trans. Pattern Anal. Mach. Intell. 35 (8) (2013) 1798–1828.

15. N. Neverova, C. Wolf, G. Taylor, F. Nebout, ModDrop: adaptive multi-modal gesture recognition, IEEE Trans. Pattern Anal. Mach. Intell. 38 (8) (2015) 1692–1706.

16. V. Kumar, S. Minz, Feature selection: a literature review, SmartC 4 (3) (2014) 211–229.

17. B. Nojavanasghari, D. Gopinath, J. Koushik, T. Baltrusaitis, L.-P. Morency, Deep multimodal fusion for persuasiveness prediction, in: Proceedings of the 18th ACM International Conference on Multimodal Interaction, 2016, pp. 284–288.

18. A. Zadeh, M. Dehmer, F. Emmert-Streib, E. Cordelli, Multimodal convolutional network for multimodal sentiment analysis, 2017, arXiv preprint arXiv:1707.02750.
[36] L.A. Hendricks, K. Burns, K. Saenko, T. Darrell, A. Rohrbach, Women also snowboard: Overcoming bias in captioning models, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 771–787.

[37] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, GRAD-CAM: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 618–626.

[38] M. Sundarrajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: International Conference on Machine Learning, PMLR, 2017, pp. 3319–3328.

[39] V. Guarrasi, N.C. D’Amico, R. Sica, L. Cordelli, P. Soda, Pareto optimization of deep networks for COVID-19 diagnosis from chest X-rays, Pattern Recognit. 121 (2022) 108242.

[40] V. Guarrasi, N.C. D’Amico, R. Sica, L. Cordelli, P. Soda, A multi-agent system to detect COVID-19 in X-ray images, in: 2021 IEEE 34th International Symposium on Computer-Based Medical Systems, CBMS, IEEE, 2021, pp. 395–400.

[41] V. Guarrasi, P. Soda, Optimized fusion of CNNs to diagnose pulmonary diseases on chest X-Rays, in: International Conference on Image Analysis and Processing, Springer, 2022, pp. 197–209.

[42] H.W. Kuhn, A.W. Tucker, Nonlinear programming, in: Traces and Emergence of Nonlinear Programming, Springer, 2014, pp. 197–215.

[43] P. Soda, A multi-objective optimisation approach for class imbalance learning, Pattern Recognit. 44 (8) (2011) 1801–1810.

[44] O. Ronneberger, P. Fischer, T. Brox, U-Net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2015, pp. 234–241.

[45] J. Shirazi, S. Katsuragawa, J. Ikene, T. Matsumoto, T. Kobayashi, K.-i. Komatsu, M. Matsui, H. Fujita, Y. Kodera, K. Doi, Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists’ detection of pulmonary nodules, Am. J. Roentgenol. 174 (1) (2000) 71–74.

[46] S. Jäger, S. Candemir, S. Fantini, V.K.I. Wáng, P.-X. Lu, G. Thoma, Two public chest X-ray datasets for computer-aided screening of pulmonary diseases, Quant. Imag. Med. Surg. 4 (6) (2014) 475.

[47] A. Krizhevsky, One weird trick for parallelizing convolutional neural networks, 2014, arXiv preprint arXiv:1404.5567.

[48] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint arXiv:1409.1556.

[49] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[50] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708.

[51] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.

[52] N. Ma, X. Zhang, H.-T. Zheng, J. Sun, ShuffleNet V2: Practical guidelines for efficient CNN architecture design, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 116–131.

[53] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, MobileNetV2: Inverted residuals and linear bottlenecks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4510–4520.

[54] M. Tan, B. Chen, R. Pang, V. Vasudevan, M. Sandler, A. Howard, Q.V. Le, MnasNet: Platform-aware neural architecture search for mobile, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 2820–2828.

[55] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 248–255.

[56] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, in: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, JMLR Workshop and Conference Proceedings, 2010, pp. 249–256.

[57] X. Glorot, A. Bordes, Y. Bengio, Deep sparse rectifier neural networks, in: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, JMLR Workshop and Conference Proceedings, 2011, pp. 315–323.

[58] A. Arcuri, G. Fraser, Parameter tuning or default values? An empirical investigation in search-based software engineering, Empir. Softw. Eng. 18 (3) (2013) 594–623.

[59] S. Rajaraman, J. Siegelman, P.O. Alderson, L.S. Folio, L.R. Folio, S.K. Antani, Iteratively pruned deep learning ensembles for COVID-19 detection in chest X-rays, IEEE Access 8 (2020) 115041–115050.

[60] D.H. Wolpert, W.G. Macready, No free lunch theorems for optimization, IEEE Trans. Evol. Comput. 1 (1) (1997) 67–82.

[61] M. Kermali, R.K. Khalas, K. Pillai, Z. Ismail, A. Harky, The role of biomarkers in diagnosis of COVID-19: A systematic review, Life Sciences 254 (2020) 117788.

[62] G.-G. Wang, S. Deb, Z. Cui, Monarch butterfly optimization, Neural Comput. Appl. 31 (7) (2019) 1995–2014.

[63] G.-G. Wang, S. Deb, L.D.S. Coelho, Earthworm optimisation algorithm: a bio-inspired metaheuristic algorithm for global optimisation problems, Int. J. Bio Inspir. Comput. 12 (1) (2018) 1–22.

[64] G.-G. Wang, S. Deb, L.D.S. Coelho, Elephant herding optimisation, in: 2015 3rd International Symposium on Computational and Business Intelligence, ISCI, IEEE, 2015, pp. 1–5.

[65] G.-G. Wang, Moth search algorithm: a bio-inspired metaheuristic algorithm for global optimization problems, Memet. Comput. 10 (2) (2018) 151–164.

[66] S. Li, H. Chen, M. Wang, A.A. Heidari, S. Mirjalili, Slim mould algorithm: A new method for stochastic optimization, Future Gener. Comput. Syst. 111 (2020) 300–323.

[67] Y. Yang, H. Chen, A.A. Heidari, A.H. Gandomi, Hunger games searches: Vi-sions, conception, implementation, deep analysis, perspectives, and towards performance shifts, Expert Syst. Appl. 177 (2021) 114864.

[68] I. Ahmadianfar, A.A. Heidari, A.H. Gandomi, RUN beyond the metaphor: An efficient optimization algorithm based on Runge Kutta method, Expert Syst. Appl. 181 (2021) 115079.

[69] J. Tu, H. Chen, M. Wang, A.H. Gandomi, The colony predation algorithm, J. Bioinsig Eng. 18 (3) (2021) 674–710.

[70] A.A. Heidari, S. Mirjalili, H. Faris, I. Aljarah, M. Mafjar, H. Chen, Harris Hawks optimization: Algorithm and applications, Future Gener. Comput. Syst. 97 (2019) 849–872.

[71] F. Saberi-Movahed, M. Mohammadifard, A. Mehrpooya, M. Rezaei-Ravari, K. Berahmand, M. Rostami, S. Karami, M. Najafzadeh, D. Hajinezhad, M. Jamshidi, et al., Decoding clinical biomarker space of COVID-19: Exploring matrix factorization-based feature selection methods, Comput. Biol. Med. 146 (2022) 105426.

[72] A. Mehrpooya, F. Saberi-Movahed, N. Azizizadeh, M. Rezaei-Ravari, F. Saberi- Movahed, M. Effekhari, I. Tavassoly, High dimensionality reduction by matrix factorization for systems pharmacology, Brief. Bioinform. 23 (1) (2022) bbab410.

[73] H.-C. Yi, Z.-H. You, D.-S. Huang, C.K. Kwoh, Graph representation learning in bioinformatics: trends, methods and applications, Brief. Bioinform. 23 (1) (2022) bbab340.

[74] W. Guo, Sparse dual graph-regularized deep nonnegative matrix factorization for image clustering, IEEE Access 9 (2021) 39926–39938.

Valerio Guarrasi received the B.C degree in Management Engineering, and the M.S. degree in Data Science from Sapienza University of Rome, Italy. He is pursuing the Ph.D. degree in Data Science at Sapienza University of Rome. His research interests include machine learning and deep learning, especially on bio-medical applications.

Prof. Paolo Soda received a Ph.D Biomedical Engineering in 2008 at Università Campus Bio-Medico (UCBM). Currently he is Full Professor of Computer Science and Computer Engineering at UCMB. His interests are in the areas of AI, machine and deep learning methods with applications to decision support systems in healthcare.