IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 25 October 2022, accepted 4 November 2022, date of publication 14 November 2022, date of current version 17 November 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3221750

RESEARCH ARTICLE

Using Artificial Intelligence for COVID-19 Detection in Blood Exams: A Comparative Analysis

ANDREA LODDO, GIACOMO MELONI, AND BARBARA PES, (Member, IEEE)
Department of Mathematics and Computer Science, University of Cagliari, 09124 Cagliari, Italy

Corresponding author: Andrea Loddo (andrea.loddo@unica.it)

This work was supported in part by the Fondazione di Sardegna through the Project ADAM (L.R. 7 agosto 2007, annualità 2018) under Grant CUP F74I19000900007, and in part by the Project ASTRID (L.R. 7 agosto 2007, annualità 2020) under Grant CUP F75F21001220007).

ABSTRACT COVID-19 is an infectious disease that was declared a pandemic by the World Health Organization (WHO) in early March 2020. Since its early development, it has challenged health systems around the world. Although more than 12 billion vaccines have been administered, at the time of writing, it has more than 623 million confirmed cases and more than 6 million deaths reported to the WHO. These numbers continue to grow, soliciting further research efforts to reduce the impacts of such a pandemic. In particular, artificial intelligence techniques have shown great potential in supporting the early diagnosis, detection, and monitoring of COVID-19 infections from disparate data sources. In this work, we aim to make a contribution to this field by analyzing a high-dimensional dataset containing blood sample data from over forty thousand individuals recognized as infected or not with COVID-19. Encompassing a wide range of methods, including traditional machine learning algorithms, dimensionality reduction techniques, and deep learning strategies, our analysis investigates the performance of different classification models, showing that accurate detection of blood infections can be obtained. In particular, an F-score of 84% was achieved by the artificial neural network model we designed for this task, with a rate of 87% correct predictions on the positive class. Furthermore, our study shows that the dimensionality of the original data, i.e. the number of features involved, can be significantly reduced to gain efficiency without compromising the final prediction performance. These results pave the way for further research in this field, confirming that artificial intelligence techniques may play an important role in supporting medical decision-making.

INDEX TERMS Covid-19 detection, artificial intelligence, machine learning, deep learning, feature selection.

I. INTRODUCTION

Covid-19 is an infectious disease caused by the Severe Acute Respiratory Syndrome CoronaVirus 2 (SARS-CoV-2) [1], declared a pandemic by the World Health Organisation (WHO) at the beginning of March 2020 [2].

The pandemic came in several waves, putting the health systems into crisis. For example, hospitals have been particularly affected by this emergency, in which intensive care cases have become a serious concern. Moreover, as of October 21, 2022, it has over 623 million confirmed cases and more than 6 million and a half deaths reported to WHO [2]. For these reasons, and even though more than 12 billion vaccines have been administered to date [2], continuous monitoring and early detection of COVID-19 positive cases remain critical to prevent the spread of the virus and to provide the most appropriate treatment for severe cases.

According to the National Institute of Allergy and Infectious Diseases (NIAID), complications brought on by a coronavirus can exhibit relevant issues that may include symptoms such as cough and breathing difficulties, fever, and kidney illnesses. In the worst cases, the disease may lead to

The associate editor coordinating the review of this manuscript and approving it for publication was Yiming Tang.

death [3]. For these reasons, governments took preventive actions and invested in research to tackle this problem.

In particular, in the field of artificial intelligence (AI), many researchers have studied and employed several machine learning (ML) and deep learning (DL) techniques to support the early diagnosis and monitoring of COVID-19. For many years, indeed, machine learning algorithms have played a crucial role in the medical field for clinical decision-making [4], [5], [6], being able to help experienced doctors, speed up the analysis process, and improve the reliability of results [7], [8], [9], [10].

More than ever, during the pandemic, these techniques have proved to be of extreme importance. An example is represented by the adoption of deep learning techniques in computer vision (CV) activities which, in some cases, even generate new data for further investigation through the support of models called generative adversarial networks (GANs) [11], [12], [13].

As witnessed by recent literature [14], [15], there is a growing demand for automated systems that can support healthcare professionals in extracting actionable knowledge from the increasing amount of digitized clinical data. There are many types of medical data, e.g., approaches based on radiography, computed tomography, and magnetic resonance imaging [16]. Similarly, the types of data made public inherent to COVID-19 are chest X-ray (CXR) images, computed tomography (CT) scans [17], [18], [19], cough waves, and many others [20], including demographic and routine clinical data.

In this context, this work aims to investigate the potential of machine learning and deep learning methods for detecting COVID-19 in blood sample data, which can potentially complement other screening and diagnostic approaches. More specifically, this work focuses on the exploration and analysis of a relatively recent dataset containing more than 40 000 instances, each described by more than 12 000 features derived from the digitization of collected blood samples.

With the purpose of building models capable of discriminating between infected and non-infected subjects, several classification methods have been applied, including Bayesian classifiers, rule-based classifiers, tree-based classifiers, instance-based classifiers, Support Vector Machines, and state-of-art ensemble methods (Random Forest and XGBoost). Several artificial neural network architectures have also been explored, leading to a final deep learning model with satisfactory performance. Furthermore, given the high dimensionality of the problem, i.e. the large number of features involved, our comparative study has explored the use of automatic feature selection techniques. They were applied in conjunction with the best-performing classification methods to provide valuable insights into which approaches may be most suitable for analyzing this type of data.

Summarizing, the main contributions of our research are the following:

1) We studied and compared the classification performance of different families of machine learning classifiers.
2) We studied and compared three deep learning classifiers recently proposed for the classification of tabular data.
3) We proposed a new artificial neural network (ANN) to handle the task at hand.
4) We investigated the extent to which feature selection can be beneficial with respect to the top-performing machine and deep learning algorithms, i.e. Random Forest and our proposed ANN.
5) We proposed an extensive comparative analysis in a domain that has not yet been fully explored and an effective pipeline to solve the task at hand.

Encompassing a large variety of methods, such a broad experimental investigation can provide valuable hints to researchers and health professionals in this field, paving the way for further, more in-depth research.

The rest of the manuscript is structured as follows. Section II provides a general review of the use of artificial intelligence techniques for COVID-19 detection. Section III presents the considered dataset and gives a brief description of the machine learning and deep learning techniques and the feature selection methods used in this work. The evaluation metrics and the leading technologies adopted are also explained. Section IV shows the experiments conducted regarding both machine learning and deep learning approaches, used alone as well as in conjunction with different feature selection techniques, with a discussion of the main findings. Finally, Section V provides final considerations on this work and outlines future research directions.

II. RELATED WORK

Over the past two years, a great deal of research has been conducted related to the diagnosis and detection of COVID-19 infections. Below we provide a general overview of previous work that, from different perspectives, sought to contribute to the battle against COVID-19 by exploiting artificial intelligence techniques.

Among the papers summarizing relevant contributions in the field, Bhattacharya et al. [21] describe several applications of deep learning in the context of COVID-19 study and analysis, including outbreak prediction, monitoring the virus spread, diagnosis and treatment, vaccine development, and drug testing.

In the work of Rasheed et al. [20], the authors illustrate how various deep learning techniques have been applied in the field of computer vision, focusing on the analysis of X-ray images. They discuss the role of AI from three perspectives: analysis, prognosis, and case tracking of COVID-19.

Another example is the work of Shorten et al. [22]. The authors studied the main applications of artificial intelligence algorithms to deal with the pandemic. They analyzed DL applications to natural language processing (NLP), life sciences, computer vision, and epidemiology, explaining how the availability of big data affects both the construction and application of learning models.

Although the research is still evolving, several works have reported significant achievements in this field. In particular, the automatic classification of COVID-19 has gained wide attention from researchers involved in the computer vision domain [23], [24], [25], [26], [27], [28], mainly thanks to the availability of imaging data like CXR or CT scans [16], [20], and even in low-end environments [29], [30].

Several works [31], [32], [33], [34], [35], [36], [37], [38], [39] have also focused on datasets containing different types of routine clinical information, including data extracted from blood tests [40], [41], [42], [43]. These datasets, often acquired under emergency conditions, are highly varied in terms of the features considered as well as the specific targets of the analysis. In some cases, the focus was on the most influential hematological features for the identification of COVID-19 positive patients [31], [32], [33], [34], [35], [36], [38]. Other works concentrated on early detection models to distinguish hospital admissions due to COVID-19 and possible entry into emergency department [37], or to distinguish between COVID-19 and influenza [39].

Within this frame of reference, several traditional machine learning algorithms were used, from Decision Trees and Random Forest [31], [32], [35] to Bayes Network [33]. Also, SVM-based strategies [36], eXtreme Gradient Boosting [37], [38], [39], and hierarchical classification systems [34] were proposed.

Some approaches have also investigated hybrid methods based on integrating clinical data with features extracted from CXR images, either handcrafted or automatically learned by convolutional neural networks [44]. The reported experimentation, conducted on patients admitted to Italian hospitals during the first wave of the pandemic, aimed at devising reliable tools for the identification of patients at risk of severe outcomes, like intensive care or death. Despite the inherent difficulty of such a complex task, the authors provided a baseline performance reference to foster further research in this direction.

Overall, the studies reported in the literature point out that the problem of automatic detection of COVID-19 from any data source is quite a difficult task. Typically, methods in the computer vision domain use the ability to infer information from imaging tools, often leading to very high performance for specific groups of patients. However, they may not be suitable for every type of COVID-19-related diagnostic scenario. On the other hand, methods based on routine clinical data, including blood sample data, may have broader applicability for larger groups of patients and can be potentially suitable for large-scale (and low-cost) screenings. Such kind of data, however, is often acquired in less controlled and heterogeneous settings, and no clear guidelines are available on the best features to consider for the analysis. In general, no single artificial intelligence approach can be optimal for each type of COVID-19-related task, motivating the exploration of different approaches that may be complementary to each other.

In this context, our work focuses on a public dataset much less explored than others but still very interesting for the considerable amount of data collected through large-scale blood tests involving more than forty thousand people. As will be presented in Section III-A, it is a challenging benchmark provided with a high number of features deriving from the digitization of the collected blood samples. Such high dimensionality makes it particularly difficult to induce accurate detection models. In addition, it does not allow direct comparison with approaches taken in previous work based on blood data.

The study that used data most similar to those employed in our work was proposed by Ribeiro et al. [45]. Indeed, their experiments were based on digitized blood samples but with lower dimensionality than the ones considered here. Specifically, the authors proposed a multilayer perceptron (MLP) with a hidden layer of 450 neurons to devise a diagnostic system with high sensitivity and specificity. Our experiments, encompassing an extensive range of learning methods, confirm the suitability of artificial neural network models in this task, as discussed below.

III. MATERIALS AND METHODS

This section presents all the materials and methods involved in our comparative analysis, including the high-dimensional dataset containing the digitized blood samples (subsection III-A), the artificial intelligence approaches adopted (subsection III-B), the evaluation metrics employed for the experimental evaluation (subsection III-C) and the chosen implementation setup (subsection III-D). Noteworthy, our study encompasses a wide range of classification techniques, which have been used both alone as well as in conjunction with different feature selection methods in order to investigate the extent to which the final classification performance varies in dependence on the data dimensionality.

A. DATASET

The employed dataset contains data extracted from blood samples collected through the blood scanner represented in Fig. 1.[1] It was released by Hilab,[2] a laboratory company from Brazil, which has thousands of blood scanner points distributed throughout the country, mostly in hospitals and pharmacies.

Given the technology of the equipment, where blood samples are digitized, and the high number of exams, enough data has become available to build a significant benchmark for machine learning and deep learning tasks. Such a benchmark has been recently used for a competition entitled *COVID19 Detection in Blood Exams*.[3]

[1] https://hilab.com.br/competition

[2] https://hilab.com.br

[3] https://www.ijcnn.org/competition-ijcnn-2021

FIGURE 1. Hilab's blood sample scanner. Picture taken from Hilab's website (link in the footnote).

TABLE 1. Data subdivision.

Fragment Id	# Instances
1	5 276
2	5 450
3	6 012
4	5 032
5	6 190
6	6 436
7	5 648
Total	40 044

It is publicly available and consists of 40 044 instances uniformly distributed into two classes representing samples positive for COVID-19 or not, as labeled by expert biomedicians. Each sample is described by 12 210 numerical features, which make the classification problem very high-dimensional and challenging.

The releasers split the dataset into seven different fragments by a stratified sampling procedure. Each one is composed of approximately 5 500 instances. Table 1 shows the number of instances for each data fragment.

B. CLASSIFICATION AND SELECTION METHODS

We employed machine and deep learning approaches to induce classification models from the considered COVID-19 detection benchmark. Furthermore, as mentioned earlier, we explored the use of different feature selection methods given the high dimensionality of the data at hand. A brief description of the methods adopted is provided below.

1) MACHINE LEARNING METHODS

For our comparative study, we exploited the following machine learning methods as representatives of different families of classifiers:

- Bayesian Network (BayesNet);
- Naïve Bayes (NB);
- Support Vector Machine (SVM);
- k-Nearest Neighbor (k-NN);
- Ripper (JRip);
- One Rule (OneR);
- Decision Tree (J48);
- Random Forest (RF);
- eXtreme Gradient Boosting (XGBoost).

Bayesian Networks are probabilistic models that represent, in the form of a *directed acyclic graph* (DAG), the conditional dependence relationships among the variables of the problem at hand (namely, in our context, the target class, and the features). The probabilistic parameters are encoded in a set of tables, one for each node of the graph, in the form of local conditional distributions of a node (variable) given its parent nodes. Once the DAG structure and the probability values have been induced from a training set of labeled examples, new instances can be classified by properly computing the posterior probability of each class value [46].

In the family of Bayesian Network classifiers, a straightforward yet effective approach is the Naive Bayes method which assumes conditional independence among the problem features, given the value of the target class. Despite this strong assumption, the Naïve Bayes approach has shown to be competitive across different classification tasks [47].

Support Vector Machines are state-of-the-art classifiers that can effectively model different types of decision boundaries and are known to scale well to high-dimensional feature spaces. In particular, the linear SVM approach involves searching for an optimal hyperplane function that maximizes the width of the margin between the classes [48]. The soft margin formulation and the so-called kernel trick allow for extending the approach to non-linearly separable problems.

The k-NN algorithm is a popular classification method in the family of instance-based learners [49] that assigns the class to unknown instances based on their similarity to the training records. Specifically, given a new instance to classify, the algorithm finds the k training records closest to it (namely, its k nearest neighbors) and makes a prediction based on a majority voting decision. A common variant is to weight the k nearest neighbors based on their distance from the unknown instance, giving higher weights to the closest neighbors [46].

Repeated Incremental Pruning to Produce Error Reduction (RIPPER) is a rule-based classifier that relies on a *sequential covering* approach [50] to induce an ordered list of prediction rules. Each rule is built greedily, starting with an empty rule antecedent and repeatedly adding conjuncts to maximize the FOIL's gain measure. The resulting rules are then refined using an incremental reduced error pruning technique. More in detail, a validation set is used to estimate the predictive performance of each rule based on a metric that is monotonically related to the rule's accuracy. Pruning is done starting from the last conjunct added to the rule: the conjunct is removed if the performance metric improves after pruning. This style of pruning has proven to be quite effective in raising predictive accuracy in noisy domains.

In the family of rule-based classifiers, One Rule is another well-known approach [51]. Basically, the algorithm constructs a rule by considering the most frequent class for each input feature's value (in the case of numerical features, they are properly discretized). Therefore, each rule is simply a set of feature values bound to their majority class. The rule with the lowest training error is finally used for prediction.

Among the tree-based classifiers, we considered the J48 algorithm, which builds a decision tree model according to the approach originally proposed by Quinlan [52]. At each node of the tree, the attribute with the highest information gain ratio is used to split the data into purer subsets. In order to reduce the risk of overfitting, the size of the tree is controlled by a post-pruning strategy based on a pessimistic error estimate made on the training data itself.

Finally, two ensemble methods were used, i.e. Random Forest and XGBoost. The Random Forest classifier relies on multiple decision trees built from different bootstrap samples of the training data [53]. In order to introduce as much diversity as possible among the ensemble components, each tree is built by selecting, for each internal node, the best splitting attribute among a set of candidate features chosen at random. Such an approach has shown a robust behavior in high dimensional spaces and, compared to other ensemble approaches, turns out to be computationally more efficient.

XGBoost [54] is an extensible gradient boosting tree algorithm that belongs to the Gradient Boosted Decision Trees (GBDT) library, introduced by Grari et al. [55]. As an ensemble grouping model, in XGBoost, new models are created from the residuals of previous models and combined to obtain the final prediction. When new models are added, a gradient descent algorithm minimizes the loss. Therefore, each tree learns from its predecessors and updates the residual errors, minimizing the errors from the previous tree.

2) FEATURE SELECTION METHODS

Feature selection, also known as variable selection or attribute selection, is a widely employed technique for reducing the original data dimensionality [56]. It involves selecting the most relevant features for the task at hand with the aim of improving the efficiency and the understandability of the induced models without degrading their performance significantly. The literature contains several approaches to formalize the concept of feature relevance and quantify the degree of relevance [57]. Nevertheless, there are no clear and standard guidelines to follow for a specific problem [58].

When used in the context of classification tasks, feature selection methods are usually categorized into three groups [59], [60]: i) *filters*, which assess the degree of correlation among the features and the target class by only relying on the intrinsic data characteristics, without interacting with the classification algorithm that will be used in the construction of the final model; ii) *wrappers*, which use a specific classifier to evaluate different candidate subsets of features (built through proper search strategies, e.g., a greedy stepwise search or an evolutionary search) and choose the one that leads to the best performance; iii) *embedded* approaches, which are based on the intrinsic ability of some classification algorithms to assign weights to the features without requiring a systematic comparison among different candidate subsets.

Due to their lower computational cost, filters are the primary choice in very high-dimensional problems, such as the one considered here. Specifically, in this work, we adopted a ranking-based selection approach, in which the features are ordered from the most important to the least important based on the strength of their correlation with the target class. Only a predefined number of highly ranked features is then used for model induction, as discussed in Section IV.

In particular, the following ranking methods, widely employed in different application domains [61], including the analysis of high-dimensional biomedical data [62], [63], [64], [65], were chosen for our experiments:

- Pearson's Correlation (Corr);
- Information Gain (InfoG);
- Gain Ratio (GainR);
- Symmetrical Uncertainty (SU);
- Mutual Information (MI).

Pearson's Correlation is a well-known criterion to evaluate the linear correlation between two variables [66]. In the context of feature selection, it assesses the worth of each attribute by evaluating the extent to which its values are linearly correlated with the class: the higher the correlation, the more relevant the attribute is for the classification task at hand.

Information Gain, Gain Ratio, and Symmetrical Uncertainty rely on the information-theoretical concept of entropy [46]. Specifically, InfoG computes a weight for each feature by measuring the extent to which the entropy of the class decreases when the value of that feature is known. GainR and SU adopt a similar approach but introduce proper normalization factors to compensate for the InfoG's bias toward features with more values.

In turn, Mutual Information is an entropic criterion to measure the degree of dependency between two variables. The specific implementation here adopted relies on a nonparametric approach based on entropy estimation from k-nearest neighbors' distances as described in [67] and [68].

3) DEEP LEARNING METHODS

Deep learning is a branch of machine learning that focuses on Artificial Neural Networks (ANNs), i.e. complex computational systems that attempt to emulate biological neural systems and employ this metaphor to learn from data.

An artificial neural network comprises a set of layers, each consisting of a collection of processing units called nodes or neurons, which are connected to each other via properly weighted directed links. Each neuron includes an activation function that determines the node's output based on the inputs received through the incoming links. The weights of the links (i.e. the network parameters) represent a fundamental aspect as the system's predictive ability depends on them.

ANN systems provide a powerful way of representing features at different levels of abstraction. In fact, at the various layers of the network, more complex features are defined starting from the raw attributes of the input dataset [66], [69]. In contrast to ''shallow'' networks that involve only a small number of hidden layers, deep neural networks are characterized by multiple layers, i.e. multiple levels of abstraction, with the ability to model very complex decision boundaries.

In order to train such complex models, adequate computational resources and advanced algorithmic procedures are required due to various factors that come into play. In particular, regularization methods play a crucial role in reducing the risk of overfitting. Further, depending on the data characteristics, proper architectural solutions need to be adopted [70]. Successful applications of such a computational paradigm are increasingly reported in the literature, across different real-world domains [71], [72], [73], [74], [75], including COVID-19 detection [25], [26], [27], [28], [44], [76].

In this work, we explored different network models to investigate the potential of deep learning in the diagnostic task at hand. The specific solution adopted, with its design choices and settings, is detailed in Section IV. Basically, it involves several intermediate layers, and the dimensionality of the input dataset is gradually reduced. Such a solution was also compared with existing state-of-the-art deep learning methods for tabular data. More specifically, the comparison included the following algorithms:

- TabNet;
- Neural Oblivious Decision Ensembles (NODE);
- 1D Convolutional Neural Network (1D-CNN).

TabNet [77] is a transformer-based model for tabular data. It comprises multiple subnetworks processed sequentially and hierarchically, like a decision tree. In particular, each subnetwork corresponds to a decision stage and receives the current batch of data as input. Then, TabNet aggregates the results of all decision phases to obtain the final prediction. TabNet first applies a sparse feature mask in each decision phase to perform a feature selection.

Instead, NODE is considered a fully differentiable model. Therefore, it permits end-to-end deep learning for training and inference employing gradient descent optimizers. Proposed by Popov et al. [78], NODE is an ensemble of differentiable oblivious decision trees [79] and uses the same splitting function for all nodes on the same level. Based on decision tree ensembles, no preprocessing or data transformation is needed.

Convolutional Neural Networks (CNNs) are rarely used on tabular data because the feature ordering has no locality characteristics. Nevertheless, a method based on a 1D convolutional neural network recently achieved the best single model performance in a Kaggle competition with tabular data [80]. More precisely, the main idea is to take advantage of CNN's property to extract features. Therefore, a fully connected layer creates a large set of features with locality characteristics, followed by multiple 1D convolutional layers.

C. EVALUATION METRICS

The metrics taken into account for the final evaluation are the following:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}; \tag{1}$$

$$Precision = \frac{TP}{TP + FP}; \tag{2}$$

$$TPR\ or\ Recall = \frac{TP}{TP + FN}; \tag{3}$$

$$FPR = \frac{FP}{FP + TN}; \tag{4}$$

$$F\_score = \frac{2 \cdot precision \cdot recall}{precision + recall}. \tag{5}$$

where TP, TN, FP, FN represent true positives, true negatives, false positives, and false negatives, respectively.

More in detail, the accuracy indicates the overall percentage of correctly classified records, as shown in Eq. (1). The precision represents the fraction of correct predictions among all instances assigned to the positive class; clearly, it depends on the number of false positives and is maximum when there are no false positives (Eq. (2)). Instead, the True Positive Rate (TPR), also known as recall, is the fraction of positive instances correctly classified as positive; it depends on the number of false negatives, with the maximum reached when there are no false negatives (Eq. (3)). A proper trade-off between precision and recall is provided by the F-score (or F-measure), which is defined as the harmonic mean of precision and recall (Eq. (5)) and takes both false positives and false negatives into account.

Another common metric is the Area Under the ROC Curve (AUC), which is a valuable criterion for comparing different classifiers [66]. Basically, the ROC (Receiver Operating Characteristic) curve is a graph that plots the True Positive Rate (TPR, Eq. (3)) against the False Positive Rate (FPR, Eq. (4)) at different probability thresholds for the positive class. Lowering the probability threshold classifies more items as positive, thus increasing both true and false positives. The area under the ROC curve provides a single score to summarize the classifier's performance on a given domain.

D. TECHNOLOGIES AND SETUP

All the experiments have been conducted on the same machine with the following configuration: Intel(R) Xeon(R) Gold 6136 CPU @ 3.00GHz CPU and Tesla P6 16 GB GPU.

Moreover, we used:

- The Weka machine learning library [46], which contains a variety of functions for classification, including the different machine learning methods described in Section III-B. It also provides functions for data preprocessing, including various feature extraction and feature selection techniques, which have been used in this work to reduce the dimensionality of the considered benchmark.
- Keras [81], a Python library that provides extensive support for deep learning; it was exploited to compare different artificial neural network architectures and implement our final deep learning model.
- Scikit-learn [82], a Python library that supports machine learning investigations and contains some feature selection techniques, particularly the MI implementation adopted in this work.

TABLE 2. Performance of the considered ML techniques. The table reports the accuracy, F-score, and AUC obtained using different data fragments as training/test sets. Values in bold are the best obtained.

	Accuracy				F-score				AUC			
Algorithm	1 & 3	3 & 1	2 & 4	4 & 2	1 & 3	3 & 1	2 & 4	4 & 2	1 & 3	3 & 1	2 & 4	4 & 2
BayesNet	0.569	0.599	0.586	0.592	0.569	0.587	0.584	0.586	0.578	0.634	0.599	0.630
NaiveBayes	0.558	0.545	0.534	0.552	0.533	0.495	0.504	0.504	0.559	0.545	0.535	0.553
Linear SVM	0.587	0.536	0.570	0.570	0.544	0.441	0.488	0.489	0.587	0.530	0.570	0.567
k-NN	0.582	0.601	0.592	0.608	0.577	0.594	0.586	0.603	0.610	0.629	0.629	0.639
k-NN (weighted)	0.583	0.601	0.592	0.607	0.577	0.595	0.586	0.602	0.614	0.631	0.636	0.642
JRIP	0.583	0.586	0.604	0.601	0.582	0.585	0.603	0.600	0.596	0.602	0.612	0.615
OneR	0.512	0.518	0.536	0.534	0.518	0.516	0.535	0.527	0.519	0.518	0.536	0.534
J48	0.598	0.581	0.586	0.585	0.597	0.581	0.586	0.585	0.606	0.590	0.583	0.594
Random Forest	0.667	**0.700**	**0.685**	**0.687**	0.667	**0.700**	**0.684**	**0.686**	**0.734**	**0.773**	**0.757**	**0.761**
XGBoost	**0.680**	0.694	0.681	0.679	**0.679**	0.694	0.681	0.678	0.679	0.694	0.681	0.679

TABLE 3. Mean values, and corresponding standard deviation, of the accuracy, F-score, and AUC obtained using different data fragments as training/test sets. Values in bold are the best obtained.

	Accuracy		F-score		AUC	
Algorithms	Mean	σ	Mean	σ	Mean	σ
BayesNet	0.587	0.013	0.582	0.008	0.610	0.027
NaiveBayes	0.547	0.010	0.509	0.017	0.548	0.010
Linear SVM	0.566	0.021	0.491	0.042	0.564	0.024
k-NN	0.596	0.011	0.590	0.011	0.627	0.012
k-NN (weighted)	0.596	0.011	0.590	0.011	0.631	0.012
JRIP	0.594	0.011	0.593	0.011	0.606	0.009
OneR	0.525	0.012	0.524	0.009	0.527	0.010
J48	0.588	0.007	0.587	0.007	0.593	0.010
Random Forest	**0.685**	0.014	**0.684**	0.014	**0.756**	0.016
XGBoost	0.684	0.007	0.683	0.007	0.683	0.007

IV. EXPERIMENTS AND RESULTS

This section presents a summary of the experimental results obtained. First, we present an investigation with different ML techniques in Section IV-A. In Section IV-B, we then explore deep learning techniques by introducing a custom artificial neural network architecture designed and implemented for this task. Finally, we discuss the results of our comparative study.

A. MACHINE LEARNING APPROACH

Here, we provide details of the experiments conducted with the machine learning approach. In particular, we have divided the experimental investigation into two main phases:

1) A preliminary analysis, in which we exploited some fragments of the dataset (see details on the data subdivision in Table 1). The results of this phase are reported in Table 2 and Table 3.
2) A detailed analysis of the entire dataset, the results of which are presented in Fig. 2 and Fig. 3. In particular, the behavior of the best performing classifier was studied in conjunction with different feature selection approaches by introducing multiple levels of dimensionality reduction.

In the preliminary analysis, we ran several tests with the considered machine learning algorithms using only the first four fragments of the dataset. In this way, we tried to get an introductory view of the data to see whether some subdivisions could produce better performance or, in general, whether any of them had more meaningful information than others. Specifically, the following configurations were considered:

- training set: fragment 1, test set: fragment 3;
- training set: fragment 3, test set: fragment 1;
- training set: fragment 2, test set: fragment 4;
- training set: fragment 4, test set: fragment 2.

The employed classifiers, described in Section III-B, were mainly trained with their default parameters. In particular, we used a linear kernel for the SVM method, while the Random Forest classifier was implemented with 100 trees and $\log_2(n)+1$ random features. In the case of the k-NN approach, with and without instance weighting, the default value of k (i.e. the number of nearest neighbors) was changed to 5; lower values, indeed, can increase the risk of over-fitting.

On the other hand, the configuration adopted for the experiments on the entire dataset was the following:

- Training set: merge of fragments 1, 2, 3, 4, and 5 ($\approx$70%);
- Validation set: fragment 7 ($\approx$14%);
- Test set: fragment 6 ($\approx$16%).

As presented in Table 1, the composition of the dataset allowed us to take advantage of stratified sampling to divide it into training, validation, and test sets, while maintaining a balanced distribution of classes.

1) RESULTS OF THE PRELIMINARY ANALYSIS

The results of the first phase of our comparative analysis are summarized in Table 2, where the accuracy, F-score, and AUC values are reported for the different configurations considered (involving fragments 1 and 3 as well as fragments 2 and 4, as explained above). The mean and standard deviation for all three metrics have also been calculated across the different configurations, as shown in Table 3. They gave more insights into the performance and behavior of each algorithm.

As seen from the tables, the results are not satisfactory overall and express the difficulty of analyzing the considered

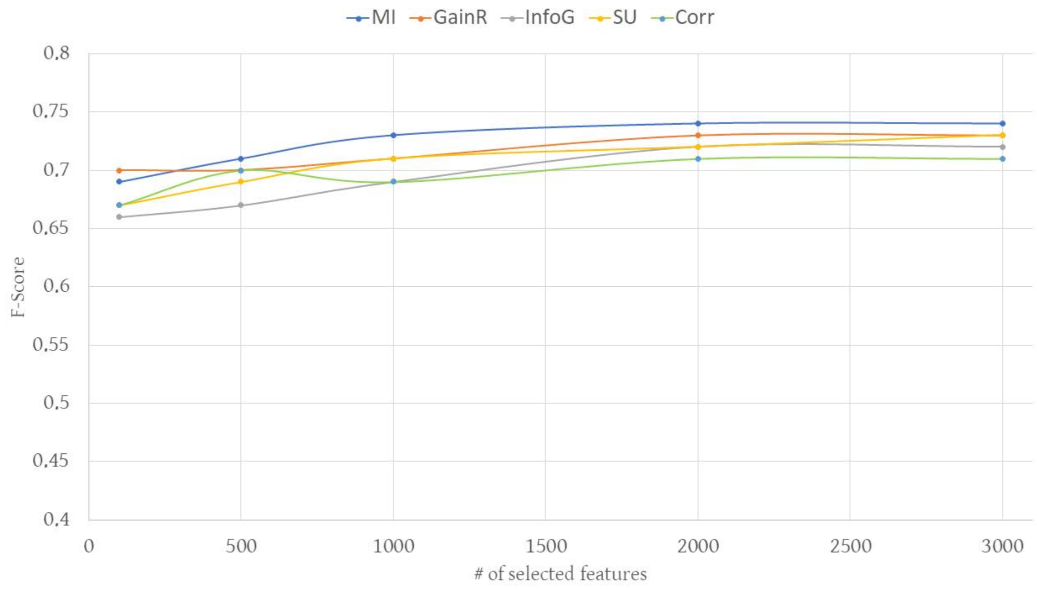


FIGURE 2. F-score performance of the Random Forest classifier in conjunction with MI, GainR, InfoG, SU and Corr selection methods, for different numbers of selected features.

high-dimensional benchmark. In each case, the Random Forest obtained the best results (emphasized in bold), with the highest values of accuracy, F-score, and AUC. This confirms the effectiveness of this ensemble approach that has proven to be a ''best of class'' learner in several tasks [83], including the analysis of high-dimensional biomedical data [62], [84], [85], [86].

These findings prompted us to focus on the Random Forest classifier for our detailed analysis of the entire dataset, as explained below.

2) RESULTS WITH FEATURE SELECTION

As previously mentioned, all the 40 044 available samples, properly divided into training, validation, and test sets, were used in this analysis phase. The experiments were carried out considering the learning method that worked best in our preliminary investigations (see Section IV-A1), i.e. the Random Forest.

More in detail, the analysis was conducted using all the original features and considering reduced feature spaces of different dimensionalities. For feature selection, the ranking techniques introduced in Section III-B2 were employed, i.e. Corr, InfoG, GainR, SU, and MI. Since each technique outputs a list in which the features appear in decreasing order of relevance, we cut this list at a proper threshold point to select the desired number of features.

Specifically, Fig. 2 shows the performance of a Random Forest model trained on increasing numbers of selected features. Different colors are used in the chart to distinguish the outcome of the different selection methods. We can see that only 100 features are sufficient to achieve an F-score value superior to 0.65. By increasing the number of selected features, the classification performance gradually improves and tends to stabilize for feature subsets containing more than 2 000 features.

TABLE 4. Hyperparameters selected for the final ANN.

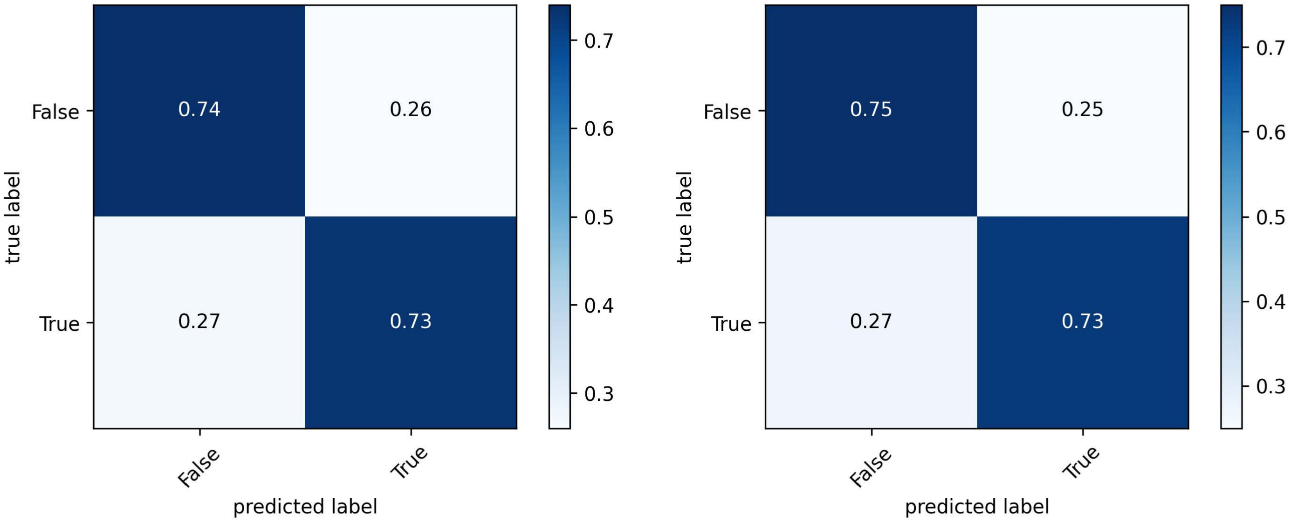
Hyperparamer	value
Epochs	1 000
Batch size	128
Activation functions	{Leaky ReLU, Sigmoid}
Loss	BCE
Optimizer	ADAM
Learning rate	1×10^{-5}

Overall, the different selection methods lead to similar results, with a slight superiority of the MI approach. With a reduced subset of 2 000 features, in fact, it leads to an F-score of almost 0.75, the same achieved on the original feature set (containing 12 210 features).

A comparison of the confusion matrices obtained with and without feature selection is provided in Fig. 3, which shows the Random Forest performance over the whole feature set (left) as well as using a reduced set of 2 000 features (right), as selected by MI. We can see that the rate of correct predictions on the positive class is the same, while the rate of correct predictions on the negative class is slightly higher using MI. Therefore, even with 84% fewer features, we achieved pretty good results, equaling those obtained with all 12 210 features.

B. DEEP LEARNING APPROACH

As additional solutions, deep learning techniques were also explored. In this regard, we conducted a large-scale preliminary investigation by implementing and testing several artificial neural network classifiers, which were trained using



(a) RF results using all features.

(b) RF results using 2 000 features.

FIGURE 3. Confusion matrices obtained with the Random Forest classifier. Fig. 3a shows the results without feature selection, while Fig. 3b shows the results over a subset of 2 000 features, as selected by the Mutual Information (MI) method.

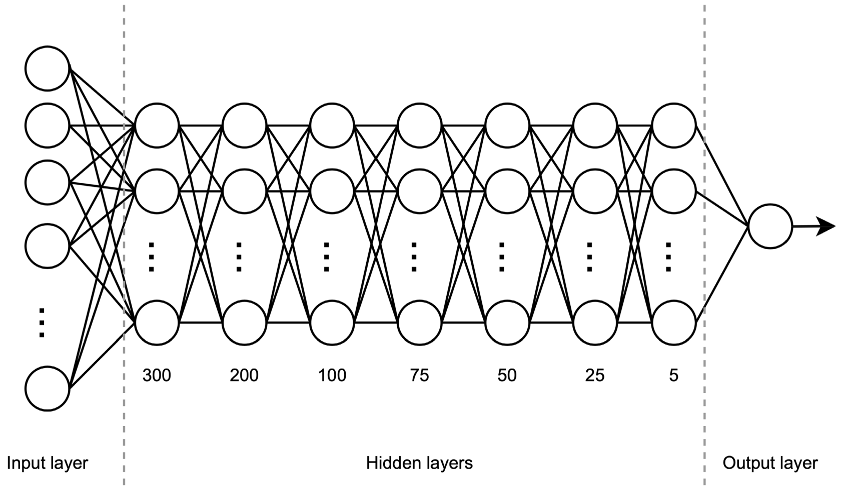


FIGURE 4. Final ANN architecture.

all the original features and on feature spaces of reduced dimensionality. The adopted configuration, the final result of these preliminary experiments, is depicted in Fig. 4. It was also compared with some state-of-the-art deep learning methods proposed for tabular data.

As can be seen, we introduced a number of intermediate layers where the dimensionality of the input layer is gradually reduced, which allowed for the extraction, through the architecture itself, of progressively fewer features (of a higher level) to be used for the final class assignment. In particular, our preliminary investigation led us to a significant initial reduction in the number of neurons, going from the input to the first hidden layer, with a more gradual dimensionality decrease through the subsequent layers. In turn, the optimal dimensionality of the input layer was also explored by introducing a preliminary feature selection step, as detailed below.

In addition to the investigation of the optimal network structure, a grid search was performed with the following hyperparameters:

- Epochs: 100, 500, 1 000, 1 500, 2 000, 3 000, 4 000, 5 000;
- Batch sizes: 16, 32, 64, 128, 256, 512, 1 024, 2 048;
- Hidden layer activation functions: ReLU, Leaky ReLU;
- Output layer activation functions: Sigmoid, Tanh;
- Loss functions: Binary Cross-Entropy, Hinge, Squared Hinge;

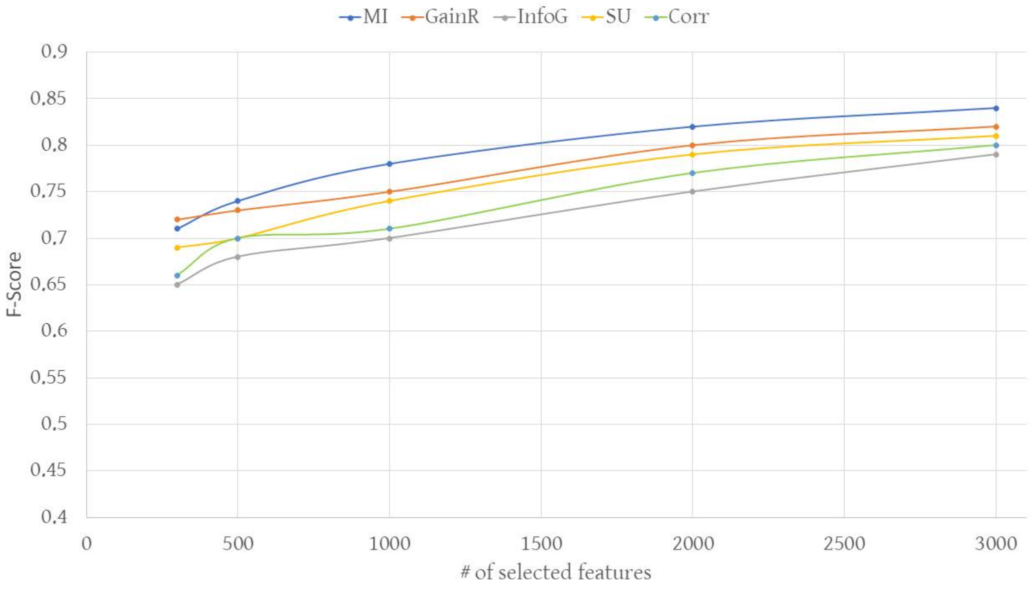


FIGURE 5. F-score performance of the proposed ANN classifier in conjunction with MI, GainR, InfoG, SU and Corr selection methods, for different numbers of selected features.

- Optimizers: Stochastic Gradient Descent (SGD), ADaptive Moment Estimation (ADAM);
- Learning rates: 1e-3, 1e-4, 1e-5, 1e-6, 1e-7.

The grid search led to the best hyperparameters summarized in Table 4. In particular, Leaky ReLU was adopted on all hidden layers, while sigmoid was only in the output layer. In addition, a dropout of 10% was added on all but the last hidden layer to avoid overfitting. It resulted in improved generalization performance. Moreover, for all models, an early stopping criterion based on the validation loss was applied. ADAM was chosen as the optimizer, with a learning rate $lr = 1e\text{-}5$ and $beta_1 = 0.9$ and $beta_2 = 0.999$, respectively. Finally, a batch size of 128 was used for 1 000 total epochs.

Our final ANN model's performance was significantly better than the one previously obtained with the machine learning approaches, including the Random Forest classifier. Indeed, using all 12 210 features of our benchmark, we got an F-score of 0.84, which is a good outcome compared to the results obtained in the competition for which the dataset was initially released (see Section III-A).

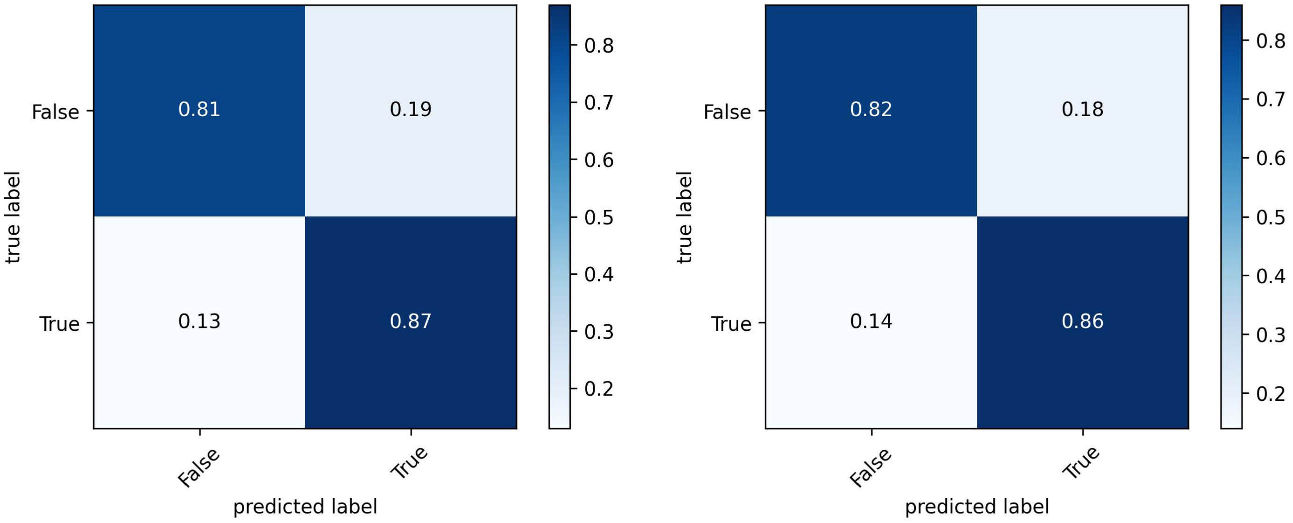
The performance obtained by the proposed ANN was also compared with other deep learning methods (see Section III-B3). The results are reported in Table 5. As can be seen, our approach outperformed the other approaches, even though the 1D-CNN model turned out to be promising as well.

We have also reported a time comparison between the methods (see Table 6) and briefly state some considerations regarding the execution time. Using the machine setup described in Section III-D, the proposed architecture needed 27 minutes to accomplish 1,000 epochs of training, while the inference time is 2.42 seconds. More specifically, Table 6 shows that our proposed architecture can obtain competitive accuracy and time-efficient (and energy-efficient) performance. In fact, it outperformed each other deep learning methods, except for TabNet, which is composed of only 26k trainable parameters.

Furthermore, as in our previous experiments, we explored the extent to which the original data dimensionality can be reduced without compromising the final classification performance. In particular, Fig. 5 shows the F-score obtained with the designed ANN classifier for different numbers of input features, as selected by the MI, GainR, InfoG, SU, and Corr ranking methods.

As can be seen, MI emerged again as the best selection technique, as also observed in the previous section for the machine learning approach. In particular, using 3 000 features selected by MI, the ANN classifier is able to reach the same performance achieved over the entire feature set, as also detailed in Fig. 6. Specifically, Fig. 6a shows the confusion matrix obtained with the proposed artificial neural network trained on all 12 210 features. In comparison, the matrix obtained by training the network on only 3 000 features (as selected by MI) is shown in Fig. 6b.

Let us finally discuss the robustness of the proposed method from two points of view: the possibility of using it as a real-time application in the healthcare domain and the results obtained. First, we consider the proposed architecture suitable for deployment in real-time systems since it consists of 3,756,386 trainable parameters. Second, we have shown that the method's performance with a reduced number of features, as selected by a proper feature selection technique, is as high as that obtained using all features. Unfortunately, as anticipated in Section II, a direct comparison of the results obtained



(a) ANN results using all features.

(b) ANN results using 3 000 features.

FIGURE 6. Confusion matrices obtained with the proposed ANN classifier. Fig. 6a shows the results without feature selection, while Fig. 6b shows the results over a subset of 3 000 features, as selected by the Mutual Information (MI) method.

TABLE 5. Accuracy, F-score, and AUC obtained using different state-of-the-art DL techniques for tabular data.

Algorithms	Accuracy	F-score	AUC
TabNet	0.802	0.786	0.808
1D-CNN	0.823	0.810	**0.868**
NODE	0.620	0.530	0.640
Our	**0.840**	**0.835**	0.841

TABLE 6. Training and inference run-time for models tested with the setup presented in Section III-D. For ease of reading, we reported the time needed to execute one epoch with the default parameters as the training time.

Method	Training	Inference
TabNet	0.82s	1.22s
1D-CNN	12.26s	6.56s
NODE	90s	18.49s
Our	1.58s	2.42s

in our work with the state of the art is not possible due to the different characteristics of the datasets used. However, taking as a reference some works based on the analysis of datasets containing hematological features, it is possible to point out that the results we have obtained are better than or in line with them. For example, Brinati et al. [31] reported an accuracy between 82% and 86% on a dataset of blood parameters they proposed [40]; Alves et al. [35] obtained an F–score of 76% on the dataset presented in [41], which consists of SARS-CoV-2 RT-PCR parameters and blood tests.

Compared with these results, we believe that the performance obtained by our proposed ANN can be deemed satisfactory, considering the high dimensionality of the data explored and the intrinsic difficulty of the related classification task.

V. CONCLUDING REMARKS AND FUTURE RESEARCH DIRECTIONS

The goal of this work was to make a contribution to the fields of machine and deep learning for the detection of COVID-19 from blood test data. To this end, we investigated and tested several approaches on a recently proposed public dataset, which proved very challenging.

First, we provided a comparative analysis of several machine learning algorithms in terms of different performance metrics. Second, deep learning techniques were also explored, leading to the proposal of an ANN architecture specifically designed for this task. Third, several feature selection techniques were investigated to reduce the dimensionality of the considered benchmark, thus allowing the construction of more efficient prediction models.

As previously discussed, Random Forest turned out to be the best-performing machine learning technique, with a rate of 73% correct predictions on the COVID-19 positive class. The proposed deep learning strategy offered a significant improvement, which outperformed the machine learning approach by correctly classifying 87% of the positive instances. Our analysis also revealed, for both Random Forest and ANN models, that the original number of features can be significantly reduced, through a preliminary feature selection step, without compromising the final classification performance. In particular, among the considered feature selection techniques, Mutual Information performed consistently better in our experiments.

Based on the different investigations conducted, we firmly believe that AI-based approaches have great potential to provide even higher results in this context. This could be achieved through a deeper analysis in multiple directions. A wider range of learning algorithms can be considered, and further architectural solutions for the deep learning approach. Furthermore, a deeper understanding of the interdependencies and correlations among the features could help improve the final classification results.

Indeed, the ranking methods here considered are the more efficient choice to reduce the data dimensionality but are not designed to capture the relationships among the features and cannot handle feature redundancy. More sophisticated selection strategies could be adopted, even relying on different selection algorithms at different stages of the selection process (e.g., initially reducing the data dimensionality through an efficient ranking approach and then further refining the search through a wrapper approach capable of optimizing the performance of a given classifier) [87], [88]. Ensemble selection methods have also recently been investigated in high-dimensional settings with promising results [89], [90].

From a broader perspective, the explored case study highlights the challenges that still need to be addressed in the context of artificial intelligence applied to COVID-19 diagnosis. In particular, the intrinsic difficulty of building high-performing classifiers from a single type of data, such as the blood sample data here considered, prompts the development of multimodal machine learning models that can process and fuse information from different data sources [34].

Finally, although artificial intelligence techniques have demonstrated remarkable performance in many diagnostic tasks, it is important to consider that medical applications require, more than others, a high level of accountability and transparency. Therefore, explanations for algorithm decisions and predictions are increasingly needed to justify their reliability and offer high interpretability for the end users [91], [92]. We also intend to explore these aspects in our future work.

REFERENCES

[1] A. Awasthi, S. Vishwas, L. Corrie, R. Kumar, R. Khursheed, J. Kaur, R. Kumar, K. R. Arya, M. Gulati, B. Kumar, S. K. Singh, N. K. Pandey, S. Wadhwa, P. Kumar, B. Kapoor, R. K. Gupta, and A. Kumar, "Outbreak of novel corona virus disease (COVID-19): Antecedence and aftermath," *Eur. J. Pharmacol.*, vol. 884, Oct. 2020, Art. no. 173381.

[2] (2022). W. H. Organization. *Who Coronavirus (COVID-19) Dashboard*. Accessed: Jul. 14, 2022. [Online]. Available: https://covid19.who.int/

[3] (2021). N. I. of Allergy and I. Diseases. *Coronaviruses*. Accessed: Apr. 13, 2022. [Online]. Available: https://www.niaid.nih.gov/diseases-conditions/coronaviruses

[4] G. D. Magoulas and A. Prentza, "Machine learning in medical applications," in *Advanced Course on Artificial Intelligence*. Torino, Italy: Springer, 1999, pp. 300–307.

[5] E. Begoli, T. Bhattacharya, and D. Kusnezov, "The need for uncertainty quantification in machine-assisted medical decision making," *Nature Mach. Intell.*, vol. 1, pp. 20–23, Jan. 2019.

[6] A. Zhong, X. Li, D. Wu, H. Ren, K. Kim, Y. Kim, V. Buch, N. Neumark, B. Bizzo, W. Y. Tak, S. Y. Park, Y. R. Lee, M. K. Kang, J. G. Park, B. S. Kim, W. J. Chung, N. Guo, I. Dayan, M. K. Kalra, and Q. Li, "Deep metric learning-based image retrieval system for chest radiograph and its clinical applications in COVID-19," *Med. Image Anal.*, vol. 70, May 2021, Art. no. 101993.

[7] S. Dash, S. K. Shakyawar, M. Sharma, and S. Kaushik, "Big data in healthcare: Management, analysis and future prospects," *J. Big Data*, vol. 6, no. 1, p. 54, Dec. 2019.

[8] X. Chen, X. Wang, K. Zhang, K.-M. Fung, T. C. Thai, K. Moore, R. S. Mannel, H. Liu, B. Zheng, and Y. Qiu, "Recent advances and clinical applications of deep learning in medical image analysis," *Med. Image Anal.*, vol. 79, Jul. 2022, Art. no. 102444.

[9] S. K. Zhou, H. N. Le, K. Luu, H. V Nguyen, and N. Ayache, "Deep reinforcement learning in medical imaging: A literature review," *Med. Image Anal.*, vol. 73, Oct. 2021, Art. no. 102193.

[10] C.-W. Wang, S.-C. Huang, Y.-C. Lee, Y.-J. Shen, S.-I. Meng, and J. L. Gaol, "Deep learning for bone marrow cell detection and classification on whole-slide images," *Med. Image Anal.*, vol. 75, Jan. 2022, Art. no. 102270.

[11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 2672–2680.

[12] T. Goel, R. Murugan, S. Mirjalili, and D. K. Chakrabartty, "Automatic screening of COVID-19 using an optimized generative adversarial network," *Cognit. Comput.*, vol. 2021, pp. 1–16, Jan. 2021.

[13] Y. Jiang, H. Chen, M. Loew, and H. Ko, "COVID-19 CT image synthesis with a conditional generative adversarial network," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 2, pp. 441–452, Feb. 2021.

[14] T. Davenport and R. Kalakota, "The potential for artificial intelligence in healthcare," *Future Healthcare J.*, vol. 6, no. 2, pp. 94–98, Jun. 2019.

[15] J. Waring, C. Lindvall, and R. Umeton, "Automated machine learning: Review of the state-of-the-art and opportunities for healthcare," *Artif. Intell. Med.*, vol. 104, Apr. 2020, Art. no. 101822.

[16] J. Ker, L. Wang, J. Rao, and T. Lim, "Deep learning applications in medical image analysis," *IEEE Access*, vol. 6, pp. 9375–9389, 2018.

[17] M. M. A. Monshi, J. Poon, V. Chung, and F. M. Monshi, "CovidXrayNet: Optimizing data augmentation and CNN hyperparameters for improved COVID-19 detection from CXR," *Comput. Biol. Med.*, vol. 133, Jun. 2021, Art. no. 104375.

[18] M. Khan, M. T. Mehran, Z. U. Haq, Z. Ullah, S. R. Naqvi, M. Ihsan, and H. Abbass, "Applications of artificial intelligence in COVID-19 pandemic: A comprehensive review," *Expert Syst. Appl.*, vol. 185, Dec. 2021, Art. no. 115695.

[19] F. A. Breve, "COVID-19 detection on chest X-ray images: A comparison of CNN architectures and ensembles," *Expert Syst. Appl.*, vol. 204, Oct. 2022, Art. no. 117549.

[20] J. Rasheed, A. Jamil, A. A. Hameed, F. Al-Turjman, and A. Rasheed, "COVID-19 in the age of artificial intelligence: A comprehensive review," *Interdiscipl. Sciences: Comput. Life Sci.*, vol. 13, no. 2, pp. 153–175, Jun. 2021.

[21] S. Bhattacharya, P. K. R. Maddikunta, Q.-V. Pham, T. R. Gadekallu, S. R. Krishnan S, C. L. Chowdhary, M. Alazab, and M. J. Piran, "Deep learning and medical image processing for coronavirus (COVID-19) pandemic: A survey," *Sustain. Cities Soc.*, vol. 65, Feb. 2021, Art. no. 102589.

[22] C. Shorten, T. M. Khoshgoftaar, and B. Furht, "Deep learning applications for COVID-19," *J. Big Data*, vol. 8, no. 1, pp. 1–54, 2021.

[23] M. Roberts et al., "Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans," *Nature Mach. Intell.*, vol. 3, pp. 199–217, Mar. 2021.

[24] D. Dong, Z. Tang, S. Wang, H. Hui, L. Gong, Y. Lu, Z. Xue, H. Liao, F. Chen, F. Yang, R. Jin, K. Wang, Z. Liu, J. Wei, W. Mu, H. Zhang, J. Jiang, J. Tian, and H. Li, "The role of imaging in the detection and management of COVID-19: A review," *IEEE Rev. Biomed. Eng.*, vol. 14, pp. 16–29, 2021.

[25] A. Loddo, F. Pili, and C. Di Ruberto, "Deep learning for COVID-19 diagnosis from CT images," *Appl. Sci.*, vol. 11, no. 17, p. 8227, Sep. 2021.

[26] A. Signoroni, M. Savardi, S. Benini, N. Adami, R. Leonardi, P. Gibellini, F. Vaccher, M. Ravanelli, A. Borghesi, R. Maroldi, and D. Farina, "BS-Net: Learning COVID-19 pneumonia severity on a large chest X-ray dataset," *Med. Image Anal.*, vol. 71, Jul. 2021, Art. no. 102046.

[27] P. Aggarwal, N. K. Mishra, B. Fatimah, P. Singh, A. Gupta, and S. D. Joshi, "COVID-19 image classification using deep learning: Advances, challenges and opportunities," *Comput. Biol. Med.*, vol. 144, May 2022, Art. no. 105350.

[28] N. Subramanian, O. Elharrouss, S. Al-Maadeed, and M. Chowdhury, "A review of deep learning-based detection methods for COVID-19," *Comput. Biol. Med.*, vol. 143, Apr. 2022, Art. no. 105233.

[29] Y. H. Bhosale and K. S. Patnaik, "IoT deployable lightweight deep learning application for COVID-19 detection with lung diseases using RaspberryPi," in *Proc. Int. Conf. IoT Blockchain Technol. (ICIBT)*, May 2022, pp. 1–6.

[30] Y. H. Bhosale, S. Zanwar, Z. Ahmed, M. Nakrani, D. Bhuyar, and U. Shinde, "Deep convolutional neural network based COVID-19 classification from radiology X-ray images for IoT enabled devices," in *Proc. 8th Int. Conf. Adv. Comput. Commun. Syst. (ICACCS)*, Mar. 2022, pp. 1398–1402.

[31] D. Brinati, A. Campagner, D. Ferrari, M. Locatelli, G. Banfi, and F. Cabitza, "Detection of COVID-19 infection from routine blood exams with machine learning: A feasibility study," *J. Med. Syst.*, vol. 44, no. 8, p. 135, Aug. 2020.

[32] A. Banerjee, S. Ray, B. Vorselaars, J. Kitson, M. Mamalakis, S. Weeks, M. Baker, and L. S. Mackenzie, "Use of machine learning and artificial intelligence to predict SARS-CoV-2 infection from full blood counts in a population," *Int. Immunopharmacol.*, vol. 86, Sep. 2020, Art. no. 106705.

[33] V. A. de Freitas Barbosa, J. C. Gomes, M. A. de Santana, J. E. D. A. Albuquerque, R. G. de Souza, R. E. de Souza, and W. P. dos Santos, "Heg. IA: An intelligent system to support diagnosis of covid-19 based on blood tests," *Res. Biomed. Eng.*, vol. 38, no. 1, pp. 99–116, 2022.

[34] M. AlJame, I. Ahmad, A. Imtiaz, and A. Mohammed, "Ensemble learning model for diagnosing COVID-19 from routine blood tests," *Informat. Med. Unlocked*, vol. 21, 2020, Art. no. 100449.

[35] M. A. Alves, G. Z. Castro, B. A. S. Oliveira, L. A. Ferreira, J. A. Ramírez, R. Silva, and F. G. Guimarães, "Explaining machine learning based diagnosis of COVID-19 from routine blood tests with decision trees and criteria graphs," *Comput. Biol. Med.*, vol. 132, May 2021, Art. no. 104335.

[36] A. T. Kouanou, T. M. Attia, C. Feudjio, A. F. Djeumo, A. N. Mouelas, M. P. Nzogang, C. T. Tchapga, and D. Tchiotsop, "An overview of supervised machine learning methods and data analysis for COVID-19 detection," *J. Healthcare Eng.*, vol. 2021, pp. 1–18, Nov. 2021.

[37] A. A. S. Soltan, S. Kouchaki, T. Zhu, D. Kiyasseh, T. Taylor, Z. B. Hussain, T. Peto, A. J. Brent, D. W. Eyre, and D. A. Clifton, "Rapid triage for COVID-19 using routine clinical data for patients attending hospital: Development and prospective validation of an artificial intelligence screening test," *Lancet Digit. Health*, vol. 3, no. 2, pp. e78–e87, Feb. 2021.

[38] M. Kukar, G. Gunčar, T. Vovko, S. Podnar, P. Černelč, M. Brvar, M. Zalaznik, M. Notar, S. Moškon, and M. Notar, "COVID-19 diagnosis by routine blood tests using machine learning," *Sci. Rep.*, vol. 11, no. 1, pp. 1–9, Dec. 2021.

[39] W. T. Li, J. Ma, N. Shende, G. Castaneda, J. Chakladar, J. C. Tsai, L. Apostol, C. O. Honda, J. Xu, L. M. Wong, T. Zhang, A. Lee, A. Gnanasekar, T. K. Honda, S. Z. Kuo, M. A. Yu, E. Y. Chang, M. R. Rajasekaran, and W. M. Ongkeko, "Using machine learning of clinical data to diagnose COVID-19: A systematic review and meta-analysis," *BMC Med. Informat. Decis. Making*, vol. 20, no. 1, pp. 1–13, Dec. 2020.

[40] D. Brinati, A. Campagner, D. Ferrari, M. Locatelli, G. Banfi, and F. and Cabitza. (Apr. 2020). *Detection of COVID-19 Infection from Routine Blood Exams with Machine Learning: A Feasibility Study*. Accessed: Jul. 14, 2022. [Online]. Available: https://zenodo.org/record/3886927#.YIluB5AzbMV

[41] (Apr. 2020). E. Hospital Israelita Albert Einstein, Sao Paulo, Brazil. *Diagnosis of COVID-19 and its Clinical Spectrum*. Accessed: Jul. 14, 2022. [Online]. Available: https://www.kaggle.com/datasets/einsteindata4u/covid19

[42] Oxford Biomedical Research Centre, Oxford, U.K. *Infections in Oxfordshire Research Database (IORD)*. Accessed: Jul. 14, 2022. [Online]. Available: https://oxfordbrc.nihr.ac.uk/research-themes-overview/antimicrobial-resistance-and-modernising-microbiology/infections-in-oxfordshire-research-database-iord/

[43] W. T. Li, J. Ma, N. Shende, G. Castaneda, J. Chakladar, and Tsai. *Classification of COVID19 and Influenza Patients*. Accessed: Jul. 14, 2022. [Online]. Available: https://github.com/yoshihiko1218/COVID19ML

[44] P. Soda, "AIforCOVID: Predicting the clinical outcomes in patients with COVID-19 applying AI to chest-X-rays. An Italian multicentre study," *Med. Image Anal.*, vol. 74, Dec. 2021, Art. no. 102216.

[45] V. H. Ribeiro, G. Steinhaus, E. Severo, J. F. Junior, L. J. Barbosa, M. Cossetin, and M. V. Figueiredo, "A system for enhancing human-level performance in COVID-19 antibody detection," in *Proc. Anais do 21st Simpósio Brasileiro de Computação Aplicada à Saúde*, Porto Alegre, RS, Brasil, 2021, pp. 224–233.

[46] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical 992 Machine Learning Tools and Techniques*, 4th ed. San Francisco, CA, USA: Morgan Kaufmann, 2016.

[47] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," *Mach. Learn.*, vol. 29, pp. 132–163, Nov. 1997.

[48] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. (2008). *Liblinear—A Library for Large Linear Classification*. The Weka classifier Works With Version 1.33 of LIBLINEAR. [Online]. Available: http://www.csie.ntu.edu.tw/c̃jlin/liblinear/

[49] D. W. Aha, D. Kibler, and M. K. Albert, "Instance-based learning algorithms," *Mach. Learn.*, vol. 6, no. 1, pp. 37–66, 1991.

[50] W. W. Cohen, "Fast effective rule induction," in *Proc. 12th Int. Conf. Mach. Learn.* San Mateo, CA, USA: Morgan Kaufmann, 1995, pp. 115–123.

[51] R. Holte, "Very simple classification rules perform well on most commonly used datasets," *Mach. Learn.*, vol. 11, pp. 63–91, Apr. 1993.

[52] R. Quinlan, *C4.5: Programs for Machine Learning*. San Mateo, CA, USA: Morgan Kaufmann, 1993.

[53] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[54] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, San Francisco, CA, USA, Aug. 2016, pp. 785–794.

[55] V. Grari, B. Ruf, S. Lamprier, and M. Detyniecki, "Fair adversarial gradient tree boosting," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Beijing, China, Nov. 2019, pp. 1060–1065.

[56] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, Jan. 2003.

[57] V. Kumar, "Feature selection: A literature review," *Smart Comput. Rev.*, vol. 4, no. 3, pp. 211–229, 2014.

[58] V. Bolón-Canedo, A. Alonso-Betanzos, L. Morán-Fernández, and B. Cancela, *Feature Selection: From Past to Future*. Cham, Switzerland: Springer, 2022, pp. 11–34.

[59] L. M. Cannas, N. Dessì, and B. Pes, "Assessing similarity of feature selection techniques in high-dimensional domains," *Pattern Recognit. Lett.*, vol. 34, no. 12, pp. 1446–1453, Sep. 2013.

[60] J. Tang, S. Alelyani, and H. Liu, "Feature selection for classification: A review," in *Data Classification: Algorithms and Applications*, C. C. Aggarwal, Ed. Boca Raton, FL, USA: CRC Press, 2014, pp. 37–64.

[61] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, *Feature Selection for High-Dimensional Data* (Artificial Intelligence: Foundations, Theory, and Algorithms). New York, NY, USA: Springer, 2015.

[62] B. Pes, "Learning from high-dimensional biomedical datasets: The issue of class imbalance," *IEEE Access*, vol. 8, pp. 13527–13540, 2020.

[63] K. K. Ghosh, S. Begum, A. Sardar, S. Adhikary, M. Ghosh, M. Kumar, and R. Sarkar, "Theoretical and empirical analysis of filter ranking methods: Experimental study on benchmark DNA microarray data," *Expert Syst. Appl.*, vol. 169, May 2021, Art. no. 114485.

[64] B. Pes and G. Lai, "Cost-sensitive learning strategies for high-dimensional and imbalanced data: A comparative study," *PeerJ Comput. Sci.*, vol. 7, p. e832, Dec. 2021.

[65] A. Bommert, T. Welchowski, M. Schmid, and J. Rahnenführer, "Benchmark of filter methods for feature selection in high-dimensional gene expression survival data," *Briefings Bioinf.*, vol. 23, no. 1, Jan. 2022.

[66] P.-N. Tan, M. Steinbach, A. Karpatne, and V. Kumar, *Introduction to Data Mining*, 2nd ed. London, U.K.: Pearson, 2018.

[67] B. C. Ross, "Mutual information between discrete and continuous data sets," *PLoS ONE*, vol. 9, no. 2, Feb. 2014, Art. no. e87357.

[68] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 69, Jun. 2004, Art. no. 066138.

[69] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016, http://www.deeplearningbook.org.

[70] F. Chollet, *Deep Learning With Python*. Shelter Island, NY, USA: Manning, Nov. 2017.

[71] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017.

[72] A. Loddo, S. Buttau, and C. Di Ruberto, "Deep learning based pipelines for Alzheimer's disease diagnosis: A comparative study and a novel deep-ensemble method," *Comput. Biol. Med.*, vol. 141, Feb. 2022, Art. no. 105032.

[73] A. Kamilaris and F. X. Prenafeta-Boldú, "Deep learning in agriculture: A survey," *Comput. Electron. Agricult.*, vol. 147, pp. 70–90, Aug. 2018.

[74] A. Loddo, M. Loddo, and C. Di Ruberto, "A novel deep learning based approach for seed image classification and retrieval," *Comput. Electron. Agricult.*, vol. 187, Aug. 2021, Art. no. 106269.

[75] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. Hoi, ''Deep learning for person re-identification: A survey and outlook,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 2872–2893, Jun. 2022.

[76] M. M. Islam, F. Karray, R. Alhajj, and J. Zeng, ''A review on deep learning techniques for the diagnosis of novel coronavirus (COVID-19),'' *IEEE Access*, vol. 9, pp. 30551–30572, 2021.

[77] S. O. Arík and T. Pfister, ''TabNet: Attentive interpretable tabular learning,'' in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, 2021, pp. 6679–6687.

[78] S. Popov, S. Morozov, and A. Babenko, ''Neural oblivious decision ensembles for deep learning on tabular data,'' 2019, *arXiv:1909.06312*.

[79] P. Langley and S. Sage, ''Oblivious decision trees and abstract cases,'' in *Proc. Workshop Case-Based Reasoning (AAAI)*, Seattle, WA, USA, 1994, pp. 113–117.

[80] (2021). Baosenguo. *Baosenguo/Kaggle-MoA-2nd-Place-Solution*. Accessed: Sep. 17, 2022. [Online]. Available: https://github.com/baosenguo/Kaggle-MoA-2nd-Place-Solution

[81] F. Chollet. (2015). *Keras*. [Online]. Available: https://keras.io

[82] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, ''Scikit-learn: Machine learning in Python,'' *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Nov. 2011.

[83] L. Rokach, ''Decision forest: Twenty years of research,'' *Inf. Fusion*, vol. 27, pp. 111–125, Jan. 2016.

[84] X. Chen and H. Ishwaran, ''Random forests for genomic data analysis,'' *Genomics*, vol. 99, no. 6, pp. 323–329, 2012.

[85] D. Chicco and L. Oneto, ''An enhanced random forests approach to predict heart failure from small imbalanced gene expression data,'' *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 18, no. 6, pp. 2759–2765, Nov. 2021.

[86] B. Pes, ''Learning from high-dimensional and class-imbalanced datasets using random forests,'' *Information*, vol. 12, no. 8, p. 286, Jul. 2021.

[87] L. M. Cannas, N. Dessì, and B. Pes, ''A hybrid model to favor the selection of high quality features in high dimensional domains,'' in *Intelligent Data Engineering and Automated Learning* (Lecture Notes in Computer Science), vol. 6936. Norwich, U.K.: Springer, Sep. 2011, pp. 228–235.

[88] N. Almugren and H. Alshamlan, ''A survey on hybrid feature selection methods in microarray gene expression data for cancer classification,'' *IEEE Access*, vol. 7, pp. 78533–78548, 2019.

[89] N. Dessì and B. Pes, ''Stability in biomarker discovery: Does ensemble feature selection really help?'' in *Current Approaches in Applied Artificial Intelligence* (Lecture Notes in Computer Science), vol. 9101. Seoul, South Korea: Springer, Jun. 2015, pp. 191–200.

[90] V. Bolón-Canedo and A. Alonso-Betanzos, ''Ensembles for feature selection: A review and future trends,'' *Inf. Fusion*, vol. 52, pp. 1–12, Dec. 2019.

[91] E. Tjoa and C. Guan, ''A survey on explainable artificial intelligence (XAI): Toward medical XAI,'' *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 11, pp. 4793–4813, 2021, doi: 10.1109/TNNLS.2020.3027314.

[92] B. H. M. van der Velden, H. J. Kuijf, K. G. A. Gilhuijs, and M. A. Viergever, ''Explainable artificial intelligence (XAI) in deep learning-based medical image analysis,'' *Med. Image Anal.*, vol. 79, Jul. 2022, Art. no. 102470.

ANDREA LODDO received the B.Sc., M.Sc., and Ph.D. degrees from the University of Cagliari, in 2012, 2014, and 2019, respectively. His Ph.D. thesis faced blood cells image analysis and classification issues to create new tools for automatic diagnosis as a support to medical analysis. He is currently an Assistant Professor with the Department of Mathematics and Computer Science, University of Cagliari. He is the author of more than 20 scientific manuscripts in peer-reviewed journals and international conference proceedings. His research interests include image analysis and processing, computer vision, pattern recognition, machine and deep learning, with particular purposes on medical tasks. Currently, he is pursuing research activities for biomedical image analysis for diagnosis support systems and videosurveillance applications.

GIACOMO MELONI received the B.Sc. and M.Sc. degrees (Hons.) in computer science from the University of Cagliari, Italy, in 2019 and 2022, respectively. During his studies, he deepened the aspects most related to machine and deep learning, two areas of great curiosity to him at the moment. His current research interest includes medical data analysis.

BARBARA PES (Member, IEEE) is currently an Associate Professor at the Department of Mathematics and Computer Science, University of Cagliari, Italy, where she taught/teaches foundations of computer science and database and data mining courses. She has participated in several research projects on web-based information systems, service-oriented architectures, data integration, high-dimensional data analysis, and bioinformatics. She is the author/coauthor of more than 80 papers published in international conferences, books, and journals. Her current research interests include the fields of data mining and machine learning, classification of high-dimensional data, and advanced feature selection methods.