Deep Generative Learning-Based 1-SVM Detectors for Unsupervised COVID-19 Infection Detection Using Blood Tests

Abdelkader Dairi, Fouzi Harrou, *Member, IEEE*, and Ying Sun

***Abstract*— A sample blood test has recently become an important tool to help identify false-positive/false-negative real-time reverse transcription polymerase chain reaction (rRT-PCR) tests. Importantly, this is mainly because it is an inexpensive and handy option to detect the potential COVID-19 patients. However, this test should be conducted by certified laboratories, expensive equipment, and trained personnel, and 3–4 h are needed to deliver results. Furthermore, it has relatively large false-negative rates around 15%–20%. Consequently, an alternative and more accessible solution, quicker and less costly, is needed. This article introduces flexible and unsupervised data-driven approaches to detect the COVID-19 infection based on blood test samples. In other words, we address the problem of COVID-19 infection detection using a blood test as an anomaly detection problem through an unsupervised deep hybrid model. Essentially, we amalgamate the features extraction capability of the variational autoencoder (VAE) and the detection sensitivity of the one-class support vector machine (1SVM) algorithm. Two sets of routine blood tests samples from the Albert Einstein Hospital, São Paulo, Brazil, and the San Raffaele Hospital, Milan, Italy, are used to assess the performance of the investigated deep learning models. Here, missing values have been imputed based on a random forest regressor. Compared to generative adversarial networks (GANs), deep belief network (DBN), and restricted Boltzmann machine (RBM)-based 1SVM, the traditional VAE, GAN, DBN, and RBM with softmax layer as discriminator layer, and the standalone 1SVM, the proposed VAE-based 1SVM detector offers superior discrimination performance of potential COVID-19 infections. Results also revealed that the deep learning-driven 1SVM detection approaches provide promising detection performance compared to the conventional deep learning models.**

***Index Terms*— COVID-19, deep learning, generative models, routine blood tests, unsupervised anomaly detection.**

I. Introduction

COVID-19, also called SARS COV-2, is a new virus pandemic fronted our world since the end of 2019. The virus starts to spread quickly with a high contagion rate until it becomes a global pandemic. Governments have taken several drastic measures to cope with the spread of the COVID-19 infection, including the quarantine of hundreds of millions of residents worldwide. On July 16, 2021, the World Health Organization (WHO) reported 188 655 968 confirmed cases of COVID-19, including 4 067 517 deaths. As of July 14, 2021, a total of 3 402 275 866 vaccine doses were administered. Nevertheless, a high number of asymptomatic cases are reported due to the COVID-19 symptomatology, making it challenging to discriminate between COVID-19 positive from negative individuals [1]. Much efforts have been made to mitigate and slowdown COVID-19 transmission by developing several techniques for different applications, such as wearing mask detection [2], COVID-19 spread forecasting [3], and chest X-ray diagnosis [4].

Accurate tests are essential for identifying positive COVID-19 cases and treating contaminated cases, which helps mitigate the pandemic [5]. Indeed, real-time reverse transcription polymerase chain reaction (rRT-PCR) becomes a standard for COVID-19 diagnosis [6], [7]. Even the RT-PCR test is needed to deal with the COVID-19 pandemic; it is still limited to certified laboratories, expensive equipment, and trained personnel, and 3–4 h needed to deliver results [6], [7]. Besides, this test can miss detecting fully symptomatic COVID-19 infected patients [8]. In other words, it has relatively large false-negative rates around 15%–20% [8]–[10]. Furthermore, it has relatively large false-negative rates [9], [10]. Therefore, alternative and more accessible, and accurate solutions are needed.

Developing accurate and fast procedures to identify infected people is undoubtedly essential to guarantee reliable control of COVID-19 spread. Recently, several machine learning-based strategies by routine blood tests have been introduced in the literature to mitigate the shortcomings of RT-PCR tests. For instance, Alves *et al.* [11] proposed a machine learning-based approach to deal with COVID-19 screening in routine blood tests. Results reveal that a random forest (RF) classifier achieved the best performance in identifying the confirmed cases based on supervised learning (i.e., accuracy 0.88, $F1$-score 0.76, and area under curve (AUC) 0.86). Also, they proposed a decision tree-based approach to explain the model and could be helpful for the health teams. In [12], ensemble learning-based methods have been applied to identify COVID-19 infected patients based on routine blood tests. Notably, a two-step model termed ERLX is proposed;

Manuscript received July 28, 2021; revised October 3, 2021; accepted November 8, 2021. Date of publication November 25, 2021; date of current version February 21, 2022. This work was supported by the King Abdullah University of Science and Technology (KAUST), Office of Sponsored Research (OSR), under Award OSR-2019-CRG7-3800. The Associate Editor coordinating the review process was Dr. Xiaotong Tu. *(Corresponding author: Fouzi Harrou.)*

Abdelkader Dairi is with the Université des Sciences et de la Technologie d'Oran Mohamed-Boudiaf (USTOMB), Oran 31000, Algérie, and also with the Laboratoire des Technologies de l'Environnement (LTE), Ecole Nationale Polytechnique Oran, Oran 31000, Algeria.

Fouzi Harrou and Ying Sun are with the Computer, Electrical and Mathematical Sciences and Engineering (CEMSE) Division, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia (e-mail: fouzi.harrou@kaust.edu.sa).

Digital Object Identifier 10.1109/TIM.2021.3130675

the RF, logistic regression (LR), and extra trees have been first applied, and their prediction outputs are used as input to XGBoost for improving the discrimination capability of XGBoost to identify COVID-19 cases. Results indicate the better classification performance of the ERLX compared to the other investigated classifiers. Wu *et al.* [13] proposed a COVID-19 infection detection-based RF algorithm using the blood test data through supervised learning. The study in [14] presented a COVID-19 infection diagnosis by applying machine learning algorithms to blood tests data with robustness to domain shifts. We compared several machine learning models in this study, such as self-normalizing neural networks, K-nearest neighbor (kNN), LR, SVM, RF, and XGB. They show that the XGB and RF exhibited better performance for COVID-19 diagnosis compared to the other models. It has been suggested to evaluate the model performance regularly to avoid a high misclassification rate. To this end, the model needs to be retrained after a certain time interval for exploiting newly collected samples. In [15], five supervised machine learning classifiers, including SVM, RF, kNN, LR, and Naive Bayes (NB), have been employed to discriminate infected COVID-19 cases healthy persons based on the blood test data. It has been concluded that machine learning classifiers can be particularly helpful in developing countries or countries facing increased infections. A set of machine learning models, including kNN, LR, NB, RF, extremely randomized trees, and decision tree, were evaluated in [16] and [17] to classify the infected cased with COVID-19 based on the blood test. We used a features selection approach to address the classification problem and identify patients who are either positive or negative to COVID-19. This study showed the feasibility of employing machine learning based on blood tests analysis as an inexpensive option to the rRT-PCR for detecting COVID-19 positive cases. This can be used as support for the countries having deficiencies of rRT-PCR reagents and limited specialized laboratories. Yang *et al.* [18] evaluated four supervised machine learning methods for COVID-19 infection detection; the models adopted are gradient boosting decision tree (GBDT), RF, LR, and DT. They trained these models in a supervised way to classify the COVID-19 infection cases—results show that the models reached an AUC of 0.854. The extreme gradient boosting machine is used in [19] to identify the infected cases of COVID-19; we performed a features selection and trained the proposed approach in a supervised manner. This study reported that the proposed approach reached a diagnostic accuracy similar and probably equivalent to RT-PCR and chest CT studies. In [20], COVID-19 infection detectors based on data from the peripheral blood of patients have been proposed. Specifically, several supervised machine learning models, including DT, RF, SVM, kNN, neural network, and variants of gradient boosting machines, have been applied to predict COVID-19 patient outcomes. It has been shown that this approach can be used to recognize infected patients with COVID-19 who are at high risk of mortality, which enables optimizing hospital resources for COVID-19 treatment. Banerjee *et al.* [21] proposed supervised machine learning models to identify SARS-CoV-2 positive patients from full blood counts without knowledge of symptoms or history of the individuals. To this end, RF and Lasso-based regularized generalized linear models and NN have been adopted. Of course, this approach could significantly improve the initial screening for patients since PCR-based diagnostic tools are limited. In [22], an approach for predicting COVID-19 PCR positivity is presented using blood count components and patient sex. This decision support tool exhibited an optimized sensitivity of 93%. Furthermore, Schwab *et al.* [23] considered five machine learning models (i.e., LR, NN, RF, SVM, and gradient boosting) to identify infected patients with COVID-19 based on routinely collected blood analysis data from a cohort of 5644 patients.

Note that all the above-mentioned approaches are based on shallow machine learning models, trained in a supervised learning approach. This study aims to develop unsupervised deep learning-driven approaches to detect COVID-19 infection based on blood test samples. This is the first study presenting unsupervised detectors to identify infected COVID-19 patients using blood test samples to the best of our knowledge. All the above-mentioned methods are designed based on supervised machine learning models where labeled data are required. Importantly, this study investigates the COVID-19 infection detection using a blood test as an anomaly detection problem through an unsupervised deep hybrid model. Overall, the contribution of this article is threefold.

1) This study addresses the problem of COVID-19 infection detection using blood test data as an anomaly detection problem. To this end, at first, the variational autoencoder (VAE) deep learning model is constructed using only anomaly-free data (without COVID-19 cases), and the VAE's extracted features are used as input for the one-class support vector machine (1SVM) to discriminate between the infected and noninfected COVID-19 patients. In terms of the methodology, our proposed VAE-1SVM approach is a fully unsupervised approach and different from the approach combining VAE and SVM for binary classification in [24]. Crucially, the method presented in [24] is a supervised approach, where the SVM is trained with labeled data, and training data contain both healthy and cancer samples. In training, the SVM classifier tries to find the appropriate hyperplane to separate these two classes using the labeled data. In short, this supervised approach needs labeled data to ensure a suitable classification. Note that the proposed VAE-1SVM approach is trained in an unsupervised manner without labeling the training data. To be more explicit, the 1SVM is trained in an unsupervised way using features extracted by the VAE encoder. The principal purpose of the 1SVM procedure is to discriminate infected from noninfected COVID-19 patients by building a hyperplane. Notably, the adoption of VAE as a features extractor is motivated by its capability to sufficiently extract the relevant nonlinear information hidden in blood test data without any prior assumption on data distribution. VAE combines the suitable characteristics of the variational inference (VI) and AE, allowing effective learning of important low-

dimensional and hidden features in data. The significant advantages of VAE over the traditional AE-based models [25] consists in its ability to alleviate the overfitting problem by incorporating a regulation mechanism in the training stage. More specifically, the regularization term enhances the ability of the generative models to sample data points utilizing learned data distribution represented in the latent space. In addition, we employed a 1SVM algorithm to separate normal and abnormal features because of its assumption-free and suitable ability to consider nonlinear features. Indeed, 1SVM maps the VAE's features in a higher feature space through kernel trick, which helps to solve linearly nonseparable cases. Accordingly, amalgamating the advantages of the VAE-driven feature extractor and the 1SVM-based detector will undoubtedly enhance COVID-19 infection detection based on blood test data.

2) This study compared the detection performance of the VAE-1SVM detector with seven deep learning models, including generative adversarial networks (GANs), deep belief network (DBN), and restricted Boltzmann machine (RBM)-based 1SVM, the traditional VAE, GAN, DBN, and RBM with softmax layer as discriminator layer, and the standalone 1SVM.
3) Two sets of routine blood tests samples from the Albert Einstein Hospital (AEH), São Paulo, Brazil, and the San Raffaele Hospital (SRH), Milan, Italy, are used to assess the performance of the investigated deep learning models. We employed multivariate data for the estimation of missing values based on the RF regressor. Results reveal that the proposed VAE-1SVM approach offers satisfying performance to identify potential COVID-19 infections and is consistently performed better than the other methods.

Section II presents the dataset and briefly describes an overview of VAE and the 1SVM algorithm. Section III presents the proposed unsupervised VAE-based 1SVM detector. In Section IV, we assess the performance of the developed approach using two publically datasets. Finally, in Section V, we conclude this study.

II. Data and Materials

This section briefly overviews the VAE model and the 1SVM classifier.

A. Data Description

In this study, two datasets of routine blood tests samples are used to assess the performance of the investigated deep learning models. The first set is collected from the AEH, SãoPaulo, Brazil, and the second one from SRH, Milan, Italy.

1) Dataset 1: The first routine blood tests samples, termed Dataset 1, are obtained from 5644 patients, including 559 infected patients with COVID-19, who had samples collected to perform the SARS-CoV-2 RT-PCR and additional laboratory tests in the AEH, São Paulo, Brazil [26]. These data are publically accessible on Kaggle website [26]. Note that these data have been normalized with zero mean and unit variance; the original data are not accessible. Dataset 1 contains 108 features; in this study, 18 important features are used based on their relevance in indicating COVID-19 based on reported studies in the literature [12], [21], [27], [28]. Fig. 1 shows the distribution of the 18 considered features, and the descriptive statistics of these features are listed in Table I. We can conclude from Table I that these datasets are non-Gaussian distributed.

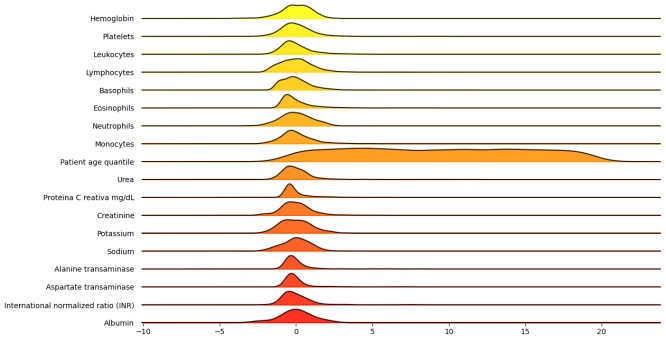

Fig. 1. Distribution of the used features in Dataset 1.

TABLE I
Summary of the Features in Dataset 1

Features	STD	Q1	Q2	Q3	Skewness	Kurtosis
Hemoglobin	0.61	-0.38	-0.01	0.41	-0.47	4.63
Platelets	0.54	-0.13	0.20	0.55	0.87	20.76
Leukocytes	0.65	-0.35	-0.01	0.18	1.38	5.93
Lymphocytes	0.54	-0.04	0.27	0.46	-0.30	4.85
Basophils	0.49	-0.14	-0.05	0.26	2.43	52.18
Eosinophils	0.47	-0.34	-0.05	0.19	2.92	36.10
Neutrophils	0.40	-0.24	-0.06	0.11	0.26	11.52
Monocytes	0.41	-0.16	-0.10	0.12	1.88	20.07
Patient age quantile	5.78	4	9	14	0.03	1.79
Urea	0.45	-0.35	-0.18	-0.02	4.46	82.37
Proteina C reativa mg/dL	0.49	-0.40	-0.25	0.04	3.64	35.11
Creatinine	0.69	-0.41	-0.05	0.24	-0.77	5.79
Potassium	0.50	-0.38	0.05	0.40	0.11	4.57
Sodium	0.41	-0.22	0.09	0.25	-0.33	15.09
Alanine transaminase	0.35	-0.30	-0.11	0.05	5.46	88.83
Aspartate transaminase	0.33	-0.32	-0.17	-0.01	6.55	113.52
International normalized ratio (INR)	0.41	-0.29	-0.09	0.07	1.89	22.00
Albumin	0.17	0.10	0.21	0.35	-0.57	12.99

2) Dataset 2: The second data are formed of three sub-datasets [15]. The first dataset consists of hematochemical values from 1624 patients at the San Raphael Hospital (OSR) collected from February to May 2020. There are 786 infected patients (48%) and 838 uninfected patients (52%). In addition, the second datasets contain 58 cases: 29 are uninfected and 29 are infected with COVID-19 collected from the Istituto Ortopedico Galeazzi (IOG), Milan, Italy, between March 5, 2020, and May 26, 2020. The third dataset, called the 2018 dataset, was obtained from blood samples gathered at the OSR in November 2018 from 54 patients. These patients are obviously uninfected from COVID-19, but 20 patients presented pneumonia-like symptoms and there employed as confounding cases. Different instruments have been utilized to collect these samples for the IOG and OSR. Thus, this composition of datasets makes Dataset 2 challenging compared to Dataset 1. Here, 11 important features have been used to detect COVID-19 infection, namely, hemoglobin (HGB),

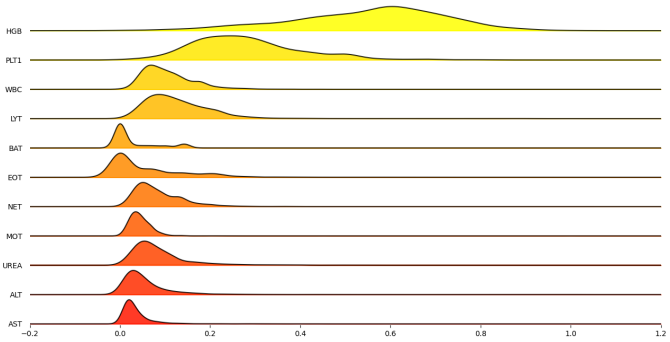

Fig. 2. Distribution of the used features in Dataset 2.

TABLE II
SUMMARY OF THE FEATURES IN DATASET 2

Features	STD	Q1	Q2	Q3	Skewness	Kurtosis
HGB	0.166	0.448	0.581	0.664	-0.476	3.032
PLT1	0.123	0.196	0.266	0.344	0.937	4.487
WBC	0.057	0.065	0.094	0.134	1.486	7.098
LYT	0.079	0.075	0.114	0.167	3.212	26.819
BAT	0.056	0.000	0.000	0.048	4.441	61.055
EOT	0.099	0.000	0.000	0.097	3.076	19.605
NET	0.055	0.047	0.072	0.113	1.925	10.505
MOT	0.035	0.028	0.041	0.055	14.036	354.584
UREA	0.075	0.046	0.068	0.105	3.736	28.110
ALT	0.073	0.022	0.040	0.067	6.137	61.829
AST	0.049	0.017	0.024	0.042	8.213	121.290

platelets (PLT1), white blood cells (WBC), lymphocytes count (LYT), basophils count (BAT), eosinophils count (EOT), neutrophils count (NET), monocytes count (MOT), UREA, alanine aminotransferase (ALT), and aspartate aminotransferase (AST). Fig. 2 and Table II show the distribution of the 11 considered features in Dataset 2. Table II shows that these datasets are non-Gaussian distributed.

B. Variational Autoencoder

It is an efficient unsupervised neural network structure that is commonly used for generative modeling [29]. The VAE introduction and concepts can be traced back to the Bayesian inference. The name of VAE comes from VI, which aims to approximate probability densities that are difficult to compute via optimization. VAE, as a stochastic generative model, has become a popular modeling technique for learning underlying distributions of the input data by reconstruction and producing new data points based on the estimated distribution. Since its first introduction in 2014, the VAE method has been found helpful in numerous applications, such as time-series forecasting [30]–[32], anomaly detection [33]–[35], and image analysis [36].

Crucially, the VAE aims to learn the probability distribution $p(\mathbf{y})$ over a multivariate variable $\mathbf{y}$. Through this, two different tasks can be accomplished. First, samples can be generated from the distribution to create new plausible values of $\mathbf{y}$. Second, to decide whether a new vector $\mathbf{y}^*$ is generated from the learned probability distribution, VAE has been intensively studied and widely employed to generate new data. The

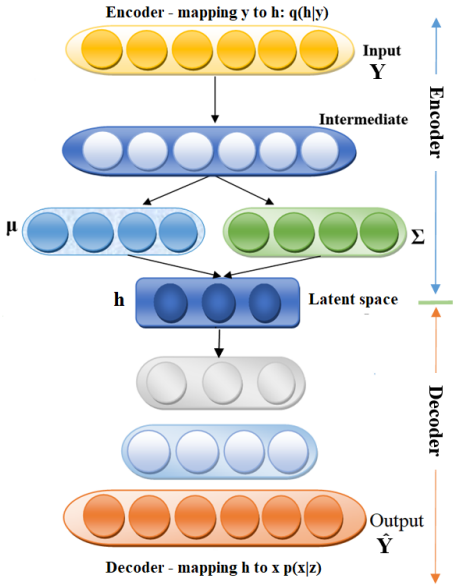

Fig. 3. VAE architecture.

primary distinction of a VAE compared to an AE consists of the encoder part and its output. Specifically, VAE is different from AE in that the VAE outputs are the parameters of the distribution generating the feature vector, while the AE is producing compressed information as a vector. We can obtain the reconstructed data by providing a randomly sampled value from the latent distribution to the decoder, making VAE able to reconstruct inputs and act as a generator, which is not the case of the AE model. Of course, the trained VAE can generate new content, which is not the case for the AE.

Overall, two neural networks are forming the VAE structure: an encoder and a decoder (Fig. 3). The purpose of the decoder is to learn a distribution (i.e., learn the parameters of the distribution) over the input data, $p(\mathbf{y}|\mathbf{h})$; it can be viewed as a Bayesian version of the principal component analysis [37]. Specifically, the encoder transforms the input data into a latent representation with a lower dimension than the original data (compacted and informative data) and stochastic (parameters of a probability distribution). The decoder plays in the opposite sense of the encoder by learning a distribution over the latent variables, $p(\mathbf{h}|\mathbf{y})$. Specifically, the decoder tries to reconstruct the input data by using the sampled data $\mathbf{h}$ from the output of the encoder.

The encoder is generally derived by a posterior approximation of $\mathbf{q}_\theta(\mathbf{z}|\mathbf{x})$, while the decoder is obtained with a likelihood $\mathbf{p}_\phi(\mathbf{x}|\mathbf{z})$, where θ and ϕ are the parameters of the encoder and the decoder, respectively.

In VAE, we get input data $\mathbf{Y}$ and aim to find the most suitable assignments of latent variables $\mathbf{h}$, which would have resulted in these data points. In other words, $\mathbf{h}$ is restricted to follow a prior distribution $p_\theta(\mathbf{h})$, usually normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$; the aim of the encoder is to learn the distribution of input data (i.e., estimate the parameters of this distribution). Mathematically, the aim is to find

$$p_\theta(\mathbf{h}|\mathbf{y}) = \frac{p_\theta(\mathbf{y}, \mathbf{h})}{p_\theta(\mathbf{y})}. \tag{1}$$

This is difficult to compute since the left-hand side in (1) and contains $p_\theta(\mathbf{y})$, which is intractable. Specifically, it can be computed via marginalizing out the latent variables

$$\begin{aligned} & p(\mathbf{y}) \\ &= \int p(\mathbf{y}|\mathbf{h})p(\mathbf{h})dh. \\ &= \iint\cdots\int p(\mathbf{y}|h_1,h_2,\ldots,h_n)p(h_1,h_2,\ldots,h_n)dh_1,\ldots,dh_n. \end{aligned} \tag{2}$$

Unfortunately, this integral is hard to compute. In VAE, this can be cast into an optimization problem and then learn the parameters of the optimization problem [38]. Specifically, this can be solved based on VI procedures by determining an approximation posterior $q_\phi(\mathbf{h}|\mathbf{y})$ [38], [39]

$$q_\phi(\mathbf{h}|\mathbf{y}) = \mathcal{N}(\boldsymbol{\mu}_\mathbf{h}, \boldsymbol{\sigma}_\mathbf{h}^2\mathbf{I}) \tag{3}$$

where $\boldsymbol{\mu}_\mathbf{h}$ and $\boldsymbol{\sigma}_\mathbf{h}$, respectively, denote the mean and standard deviation of $q_\phi(\mathbf{h}|\mathbf{y})$, which are calculated through the encoder.

Given $q_\phi(\mathbf{z}|\mathbf{x})$, the evidence lower bound (ELBO) is calculated as [38], [39]

$$\log p_\theta(\mathbf{y}) = \mathbb{E}_{q_\phi(\mathbf{h}|\mathbf{y})}\left[\log p_\theta(\mathbf{y})\right] \tag{4}$$

$$= \mathbb{E}_{q_\phi(\mathbf{h}|\mathbf{y})}\left[\log \frac{p_\theta(\mathbf{y}|\mathbf{h})p_\theta(\mathbf{h})}{p_\theta(\mathbf{h}|\mathbf{y})}\right] \tag{5}$$

$$= \mathbb{E}_{q_\phi(\mathbf{h}|\mathbf{y})}\left[\log \frac{p_\theta(\mathbf{y}|\mathbf{h})p_\theta(\mathbf{h})}{p_\theta(\mathbf{h}|\mathbf{y})}\frac{q_\phi(\mathbf{h}|\mathbf{y})}{q_\phi(\mathbf{h}|\mathbf{y})}\right] \tag{6}$$

$$\begin{aligned} &= \mathbb{E}_{q_\phi(\mathbf{h}|\mathbf{y})}\left[\log p_\theta(\mathbf{y}|\mathbf{h}) + \log p_\theta(\mathbf{h}) - \log q_\phi(\mathbf{h}|\mathbf{y})\right] \\ &\quad + D_{\mathrm{KL}}(q_\phi(\mathbf{h}|\mathbf{y})||p_\theta(\mathbf{h}|\mathbf{y})) \end{aligned} \tag{7}$$

where $D_{\mathrm{KL}}[.]$ is the Kulback–Leibler divergence of the approximate $q_\phi(\mathbf{h}|\mathbf{y})$ from the true posterior $p_\theta(\mathbf{h}|\mathbf{y})$, and the first term is the ELBO. Since the KL divergence is positive (i.e., in (7), $D_{\mathrm{KL}}(q_\phi(\mathbf{h}|\mathbf{y})||p_\theta(\mathbf{h}|\mathbf{y})) \geq 0$), then we have

$$\log p_\theta(\mathbf{y}) \geq \mathbb{E}_{q_\phi(\mathbf{h}|\mathbf{y})}\left[\log p_\theta(\mathbf{y}|\mathbf{h}) + \log p_\theta(\mathbf{h}) - \log q_\phi(\mathbf{h}|\mathbf{y})\right]. \tag{8}$$

The term on the right-hand side of (8), also called the ELBO, is, therefore, the lower bound of $\log p_\theta(\mathbf{y})$, which wants to maximize. Accordingly, to maximize $\log p_\theta(\mathbf{y})$, we can focus on maximizing the ELBO term, which is equivalent (but computationally tractable). The loss function of the VAE becomes

$$\begin{aligned} & \mathcal{L}_{\mathrm{VAE}}(\theta,\phi;\mathbf{y}) \\ &= \mathbb{E}_{q_\phi(\mathbf{h}|\mathbf{y})}\left[\log p_\theta(\mathbf{y}|\mathbf{h}) + \log p_\theta(\mathbf{h}) - \log q_\phi(\mathbf{h}|\mathbf{y})\right] \quad (9) \\ &= \mathbb{E}_{q_\phi(\mathbf{h}|\mathbf{y})}\left[\log p_\theta(\mathbf{y}|\mathbf{h}^*)\right] - D_{\mathrm{KL}}(q_\phi(\mathbf{h}|\mathbf{y})||p_\theta(\mathbf{h})). \quad (10) \end{aligned}$$

In summary, at first, the VAE model encodes input data as a distribution over the latent space (Fig. 4). Notably, the data points are encoded as a distribution over the latent space based on the VAE decoder. Then, observations from the latent space are sampled from that distribution. After that, the sampled observations are decoded, and the reconstruction error is calculated. Finally, the reconstruction error is backpropagated through the network. In the training stage, the stochastic gradient variational Bayes algorithm is usually applied to optimize the ELBO to determine the values of the encoder and decoder parameters [40], [41]. For more details on the VAE, see [38].

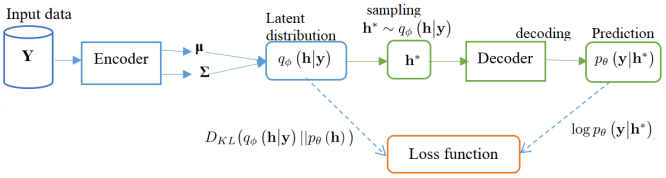

Fig. 4. Illustration of the basic concept of VAE structure.

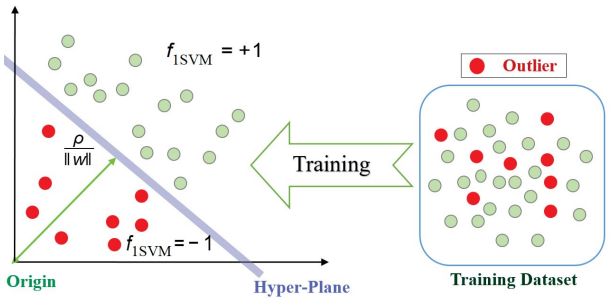

Fig. 5. Illustration of the basic concept of 1SVM procedure.

C. One Class SVM

The particularity of the 1SVM scheme resides in its training with only anomaly-free observations to learn nominal behavior in the investigated process, and no labeling is needed to construct 1SVM. Crucially, 1SVM is an unsupervised binary classifier [42]. 1SVM aims to determine a hyperplane as far as possible from the origin; mainly, the hyperplane should be as close as possible to the normal observation of the training data points in the original space [43]. Then, it is used to classify new testing data as comparable or different from the training data. 1SVM has been widely employed for anomaly detection in different applications, including photovoltaic (PV) systems monitoring [44], obstacle detection in autonomous vehicles [45], wastewater treatment plant monitoring [46]–[48], and air quality monitoring [49].

A mapping function is employed to make the indivisible samples in low-dimensional space easier to separate in a high-dimensional space. Specifically, 1SVM aims to construct a hyperplane with the maximum margin to separate the data points from the origin (Fig. 5) as formulated by the following equation:

$$\begin{aligned} & \min_{\omega\gamma\rho}\left(\frac{1}{2}\omega^T\omega - \rho + \frac{1}{\upsilon l}\sum_{i=1}^{l}\gamma_i\right) \\ & \text{s.t.}: \omega.\ \Psi(x) > \rho - \gamma \end{aligned} \tag{11}$$

where l is the size of training data, ω denotes a weight vector, $\upsilon \in (0,1]$ refers to the regularization term, and γ is the nonzero slack variable utilized for penalizing the observation

that may fall outside the decision boundary during the training stage. The term ρ refers to the margin between the origin and the mapped samples in feature space; it is also called offset. 1SVM employs a decision function $\mathcal{F}$ given in (12) that returns -1 for an anomaly and 1 for a typical data point based on the hyperplane

$$\mathcal{F}(x) = \text{sign}(\omega.\ \Psi(x) - \rho) \tag{12}$$

where Ψ is applied to map the data samples into a higher dimensional feature space. The term $(\rho/\|\omega\|)$ defines the hyperplane (see Fig. 5), it is, in fact, Euclidean distance from the origin to the support vector point, and this term must be maximized. The 1SVM quadratic optimization problem given in the following equation:

$$\min_{\omega\gamma\rho}\left(\frac{\|\omega\|^2}{2} - \rho + \frac{1}{\upsilon l}\sum_{i=1}^{l}\gamma_i\right). \tag{13}$$

Hence, during the training, the objective function encourages the maximization of the margin $(\|\omega\|^2/2) - \rho$ and minimizing the average of the slack variables γ. In our study, we adopt the Gaussian radial basis function (RBF) $\mathcal{K}$, defined in the following equation:

$$\mathcal{K}(x, x') = \langle \Psi(x), \Psi(x') \rangle = e^{(\alpha\|x-x'\|^2)} \tag{14}$$

where $\|x - x'\|^2$ is the dissimilarity measure and α denotes the kernel parameter.

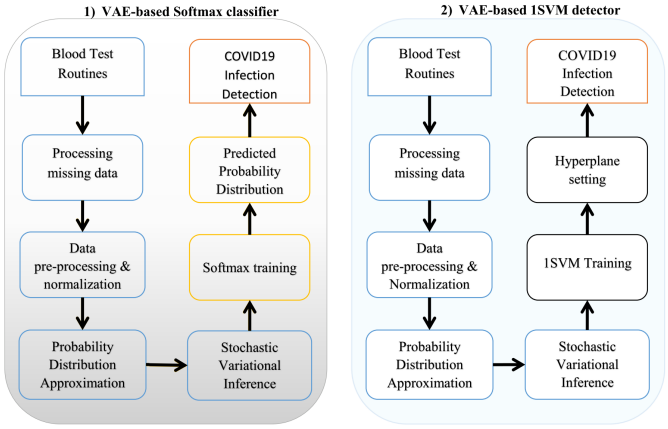

Fig. 6. Schematic representation of the proposed schemes: 1) VAE-based softmax classifier and 2) unsupervised VAE-based 1SVM detector.

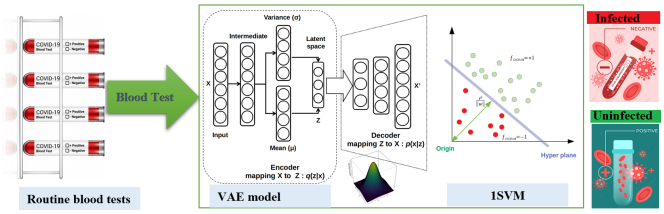

Fig. 7. Flowchart of the unsupervised VAE-based 1SVM detector.

III. Methodology

This study proposes two different frameworks for COVID-19 infection detection (Fig. 6): an unsupervised VAE-1SVM detector and VAE-Softmax classifier. The first approach is based on learning the data probability distribution of the COVID-19 blood tests data through a stochastic VI process; the constructed model is combined with a softmax classifier called the VAE-Softmax classifier. On the other hand, in the second approach, we address COVID-19 infection using blood test data as an anomaly detection problem. Specifically, we train the model using only noninfected blood test data to extract features that will feed the 1SVM to identify infected cases, and we called this approach VAE-1SVM.

However, the adopted datasets in this study suffer from the missing data problem; in order to overcome this situation, multivariate data for estimation of missing values based on RF regressor are used [50], [51]. This technique estimates the missing values by considering the other variables that improve the quality of the imputed data.

As VAE-Softmax is a relatively well-known approach, this section will present only the VAE-based 1SVM approach.

A. Unsupervised VAE-Based 1SVM Detector

The VAE-1SVM approach amalgamates the advantages of a deep generative model, namely, VAE, with a 1SVM detector (see Fig. 7). The main differences between this approach and the previous one are two folds; this approach addresses infection detection as an anomaly detection using only noninfected data construct VAE-1SVM unsupervised detector. Specifically, the VAE focuses only on noninfected data; in other words, the VIs procedure will be applied to one class of data to learn its probability distribution to approximate it through sampling from the latent space. Hence, after completing the unsupervised VAE training, the latent space will be suitable for generating data points sharing the same features of the training data. At this stage, VAE is achieving several tasks simultaneously, dimensionality reduction features extraction, and encoding a new compact data representation of the original blood test of the noninfected cases. Note that the testing dataset, which will be used to evaluate the classification performance of the proposed approach, comprises both infected and noninfected observations. The resulting latent space is used to feed the 1SVM, which is sensitive to the outlier points in the training sets. It is expected that the VAE will improve the 1SVM capability of building a hyperplane with a maximum margin to separate the data points (inliers) from the origin that separates the outliers. Here, the RBF kernel is used to estimate the probability density function (PDF) of the blood test dataset encoded by VAE. More specifically, the 1SVM objective function is responsible for deciding whether there is a COVID1-9 infection or not. To be concise, the 1SVM is applied to the features extracted by the VAE model (i.e., the output of the VAE encoder) to detect the COVID1-9 infection. At first, we train the 1SVM using the VAE discriminator output based on training data that include only positive (noninfected) cases; no labeling is used for constructing 1SVM. The training step aims to find a hyperplane as close as possible to the normal samples. Next, it is employed to assess new test samples as similar or distinct from the training samples.

TABLE III
HYPERPARAMETER SETTINGS OF RBM, DBN, GAN, VAE, AND 1-SVM

Model	Parameters
RBM	hidden units = 20 , K=5
DBN	Layer 1 : hidden units = 30 , Layer 2 : hidden units = 10, K=5
GAN	Generator [Layer 1: 18, Layer 2: 18], Discriminator [Layer 1: 18, Layer 2: 11]
VAE	Intermediate = 18 Mean=20, Variance=20, and Z=10
1-SVM	kernel=RBF , nu=0.0025 , gamma=0.01

IV. RESULTS AND DISCUSSION

This section evaluates the proposed approach model and discusses the results using Datasets 1 and 2. In this study, we conduct two experiments for each dataset. The first experiment is based on deep learning generative models (i.e., GAN, DBA, RBM, and VAE) coupled with a softmax classifier to perform the COVID-19 infection detection via classification of confirmed and clean cases. Furthermore, the second experiment is based on hybrid models built up by stacking the already used deep generative learning models with the unsupervised 1SVM detector for COVID-19 infection detection. The performance of the proposed approaches is evaluated using the standard classification metrics: true-positive rate (TPR), false-positive rate (FPR), accuracy, precision, $F1$-score, and AUC.

Note that to train the unsupervised VAE-, GAN-, RBM-, and DBN-based 1SVM detectors, we choose randomly 80% from the noninfected observation. The remaining data are used for testing and contain both infected and uninfected cases. In other words, the proposed approaches are trained based only on data from uninfected cases. However, for the VAE-, GAN-, RBM-, and DBN-based softmax classifiers, each dataset is divided into two subsets: training composed of 80% (infected and uninfected cases) and testing with 20% (infected and uninfected cases).

A. Settings

This study evaluates the performance of four deep generative models: GAN, VAE, RBM, and DBN. Optimal parameters of each model are carefully chosen (parameter tuning phase) using the training data based on the grid search procedure. Here, common settings have been used: binary cross entropy is used as loss function, Rmsprop optimizer, batch size of 50, 2000 epochs, and learning rate of 0.001. The values of the parameters of the investigated models in this work are arranged in Table III.

RBM consists of two layers, the visible layer and the hidden layer representing the latent variables representing a compact version of the learned data probability distribution. Note that RBM is part of undirected graph models trained in an unsupervised way based on the Gibbs sampling algorithm that belongs to Markov chain Monte Carlo (MCMC) techniques. Furthermore, we added the softmax layer for the classification, where the entire architecture (RBM + Softmax) is fine-tuned using supervised learning. In all RBM evaluations, we found that setting the hidden layer dimension to 20 is the best, where the visible layer dimension is obtained based on the input size of the dataset used.

The DBN architecture used in our experiment is composed of two stacked RBMs, where we adopt a greedy layerwise training approach to train the DBN model. Importantly, first, each RBM is trained in an unsupervised way, and second, a softmax layer is added to the stacked RBMs to form a deep neural network model (RBM1 + RBM2 + Softmax). Then, this new model is fine-tuned via supervised learning to adjust and optimize the learned parameter to assign a probability for each new observed cased to a given class (infected or not) with COVID-19. The following configuration has been adopted: RBM1 with 30 hidden units and RBM2 with ten hidden units.

On the other hand, the GAN model consists of two distinct deep neural networks: generative and discriminative. Thus, the configuration (hyperparameters) of the two parts needs to be determined to reach the highest classification performance.

For the GAN model, it has been based on an unsupervised manner according to the zero-sum game. The zero-sum encourages the discriminator to successfully identify real and fake samples, whereas the generator is penalized with large updates to generate samples similar (i.e., share the same features space) to the training data points. Based on grid search, we used the following settings for GAN that are adopted: generative with (Layer 1: 18, Layer 2: 18) and discriminative with (Layer 1: 18, Layer 2: 11). After completing the training, the discriminative is used with a softmax to achieve the classification task forming (discriminative + Softmax) model trained in a supervised way.

Finally, the VAE is composed of five layers (Fig. 3): input, intermediate, colorblackmean, variance, and Z. The input layer dimension depends on the dataset used (18 and 20); the intermediate layer is set to 30 hidden units, while mean, variance, and Z dimensions are set to ten hidden units. As a generative model, VAE is trained first through an unsupervised way based on the stochastic VIs approach, where a regularized iterative sampling is done that aims to sample during the training new data points with the same features as the training dataset. The VAE is combined with a softmax layer to form a deep neural network (VAE-Softmax), trained here in a supervised manner to fine-tune the models' parameters learning during the unsupervised training to deal with the classification problem, which aims to assign a probability to each class. The 1SVM is used as a standalone unsupervised classifier and is used as a substitution to the softmax with only one output (1 if the case is infected or 0 otherwise). The 1SVM has three main parameters, and we set their values based on grid search as follows: the RBF kernel, $\nu = 0.0025$, and $\gamma = 0.01$. Several kernels could be considered in designing the 1SVM scheme. Note that there is no automatic procedure for selecting the optimal kernel. Here, we evaluated the performance of the 1SVM with three kernels, namely, RBF, polynomial, and sigmoid, based on training data. Then, we used the 1SVM with RBF kernel because it achieved the highest detection performance. RBF kernel enables the 1SVM to perform as a linear or nonlinear detector.

B. Experimental Results

In this study, we attempted to detect infected patients with COVID-19 using blood test exams data. We addressed this

problem as an anomaly detection problem, and the output of the detectors will be infected or not infected based on the blood test data. Here, we evaluate the efficiency of the proposed approach that the flexibility of deep generative models with the sensitivity 1SVM model is to deal with COVID-19 infection detection. In this study, two approaches to detect COVID-19 infection are adopted through a set of considered deep generative models: VAE, RBM, DBN, and GAN. Notably, three different approaches are considered for probability distributions approximation.

1) *Finding Nash Equilibrium in Two-Player, Zero-Sum Games:* Here, the GANs can be viewed as a zero-sum game between two machine players, a generator and a discriminator, constructed for learning data distribution.
2) *MCMC:* Here, RBM and DBNs belong to MCMC methods. In such techniques, the MCMC algorithm (Gibbs sampling) or sampling from a probability distribution is used in training.
3) *VI:* The VAE is based on the VI framework, which approximates probability densities by optimization. The VI was employed in various applications and tends to be faster than traditional techniques (e.g., MCMC sampling) [40]. The essence of VI consists of positing a family of densities and then finding a family member closer to the target density [40]. The Kullback–Leibler divergence is commonly used to measure the closeness between two distributions.

Note that for RBM- and DBN-based 1SVM, the RBM and DBN are used as feature extractors, and the 1SVM is applied to the output of RBM and DBN models to uncover infected cases. Similar to VAE-1SVM, in RBM-1SVM and DBN-1SVM, the 1SVM is first trained based on samples from healthy cases, and then, it is used to uncover potential infected cases in test samples.

Table IV lists the obtained validation metrics of testing data from VAE-, GAN-, RBM-, and DBN-based softmax schemes based on blood test Dataset 1. Note that here, the data used include both infected and noninfected blood test observations. We first observe that all generative models achieve an AUC greater than 0.93. We also observe that the VAE-based approach is the best approach for identifying infected cases from noninfected ones in terms of all calculated metrics. Specifically, VAE reached an AUC of 99.85% and outperformed GAN (AUC = 99.53%) slightly, while RBM and DBN recorded AUC of 93.53% and 94.19%, respectively (Table IV). Moreover, VAE scored the highest classification performance in accuracy, precision, and $F1$-score compared to the other deep generative model (Table IV). Results in terms of FPR and TPR metrics are shown in Fig. 8. Results report the better performance of the VAE-based method by providing the lowest FPR and the highest TPR compared to the other models. The obtained results show the superiority of VAE combined with softmax to deal with infection or noninfection of COVID-19 classification task based on blood test data obtained from more than 5000 patients. This could be attributed to the effectiveness of the stochastic VI performed by the VAE to explain the most variance in the blood test data.

TABLE IV
CLASSIFICATION RESULTS OF COVID-19 BLOOD TEST USING DATASET 1

MODEL	TPR	FPR	Accuracy	Precision	F1Score	AUC
GAN	0.9960	0.0246	0.9938	0.9970	0.9965	0.9953
VAE	0.9970	0.0088	0.9965	0.9990	0.9980	0.9985
RBM	0.9795	0.2330	0.9601	0.9767	0.9781	0.9419
DBN	0.9772	0.2333	0.9548	0.9724	0.9748	0.9353

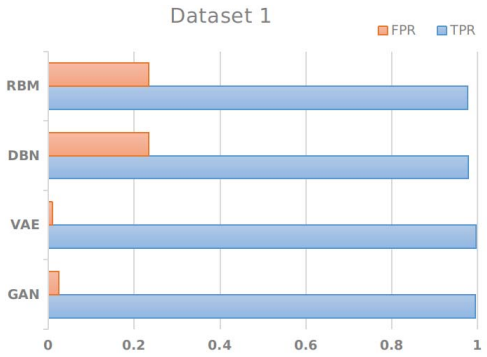

Fig. 8. Obtained FPR and TPR of GAN-, VAE-, RBM-, and DBN-based softmax methods based on Dataset 1.

TABLE V
DETECTION RESULTS OF COVID-19 BLOOD TEST USING DATASET 1

MODEL	TPR	FPR	Accuracy	Precision	F1Score	AUC	
VAE-1SVM	1.000	0.009	0.999	0.999	1.000	1.000	0.996
GAN-1SVM	1.000	0.017	0.998	0.998	1.000	0.999	0.991
RBM-1SVM	0.994	0.107	0.983	0.987	0.994	0.991	0.943
DBN-1SVM	1.000	0.034	0.996	0.996	1.000	0.998	0.983
1SVM	1.000	0.130	0.985	0.983	1.000	0.992	0.935

The detection results based on the standalone 1SVM and VAE-, GAN-, RBM-, and DBN-based 1SVM detectors are listed in Table V. It should be noted that these detectors are trained in an unsupervised manner where only the noninfected blood test data are used in the training stage. Specifically, we used 85% of the noninfected observations for training, while the testing set contains the remaining 15% noninfected observations in addition to the infected observations. Table V shows that the unsupervised detectors achieved satisfactory performances with an AUC greater than 0.943%. We also conduct the COVID-19 infection detection based on a standalone 1SVM; it achieved an acceptable performance with AUC = 93.5%. Results demonstrate that using deep generative models as features extractors followed by the 1SVM algorithm offers better detection accuracy than the standalone 1SVM. Again, in terms of all metrics calculated, the VAE-1SVM detector is the best approach in detecting COVID-19 infection by reaching an AUCof 99.6%. It can also be seen that the GAN-1SVM detector achieved high efficiency and satisfying accuracy with an AUC of 99.1% (Table V). Recall here that we address COVID-19 infection detection based on blood tests as an anomaly detection problem; the proposed approach hybrid model has demonstrated a high detection performance based on totally unsupervised learning.

Now, the efficiency of the considered methods will be verified using Dataset 2. As described above, these data are

TABLE VI
CLASSIFICATION RESULTS OF COVID-19 BLOOD TEST USING DATASET 2

Model	TPR	FPR	Accuracy	Precision	Recall	F1Score	AUC
GAN	0.802	0.191	0.805	0.724	0.809	0.764	0.891
VAE	0.849	0.176	0.839	0.778	0.824	0.800	0.900
DBN	0.764	0.191	0.782	0.688	0.809	0.743	0.881
RBM	0.792	0.162	0.810	0.722	0.838	0.776	0.890

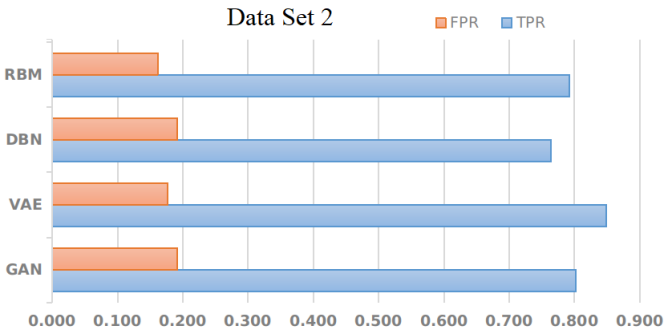

Fig. 9. Obtained FPR and TPR of GAN-, VAE-, RBM-, and DBN-based softmax methods based on Dataset 2.

the concatenation of three different subsets of data. These data are collected from different geographic locations by using different medical types of equipment. It contains fewer samples of blood tests, about 1700, compared to Dataset 1. This data heterogeneity makes the classification problem more challenging.

The binary classification results of COVID-19 infection based on VAE-, GAN-, RBM-, and DBN-based softmax schemes when applied to blood tests Dataset 2, which consists of infected and noninfected observation, are summarized in Table VI. Here, the output layer of the considered generative models consists of a softmax. It can be observed that the overall AUC obtained is around 88% and 90%, which is acceptable regarding the Dataset 2 heterogeneous contents and also the unbalanced number of cases on infected and noninfected. In this experiment part, VAE (AUC = 90%) outperforms slightly GAN (AUC = 89.1%), RBM (AUC = 89%), and DBN (AUC = 88.1%). Fig. 9 shows the FPR and TPR reported by the investigated approaches. Similar conclusions are drawn; VAE has recorded the lowest FPR 0.176 and the highest TPR 0.849.

The final experiment aims to assess the potentials of the deep generative models combined with 1SVM to detect COVID-19 blood test infection using Dataset 2. Table VII summarizes the comparative detection results of the considered models. We can see that the detection performance is considerably enhanced compared to the previews results given in Table VI. The AUC obtained by VAE was improved from 90% to 99.3%; all considered deep generative models' performances improved to an AUC greater than 96%, satisfying detection results. We can see that our VAE-1SVM detector recorded the best score with the highest detection performance (i.e., AUC = 0.993). Furthermore, in this experiment, we compared the GAN-, VAE-, RBM-, and DBN-based 1SVM detectors with the standalone 1SVM for COVID-19 infection detection (Table VII). Results show again that using VAE-1SVM improves the detection quality from AUC = 96% to AUC = 99.3% compared to the standalone 1SVM.

TABLE VII
DETECTION RESULTS OF COVID-19 BLOOD TEST USING DATASET 2

Model	TPR	FPR	Accuracy	Precision	Recall	F1Score	AUC
1SVM	0.931	0.010	0.966	0.985	0.931	0.957	0.960
VAE-1SVM	0.986	0.000	0.994	1.000	0.986	0.993	0.993
GAN-1SVM	0.944	0.010	0.971	0.985	0.944	0.964	0.967
RBM-1SVM	0.971	0.000	0.989	1.000	0.971	0.986	0.986
DBN-1SVM	0.944	0.000	0.977	1.000	0.944	0.971	0.972

Overall, first, this study shows the superiority of VAE combined with softmax classifier to deal with COVID-19 infection detection through a classification approach for the two considered datasets. In short, results reveal the superior performance of the VAE-Softmax classifier compared to GAN-, RBM-, and DBN-based softmax classifiers. This could be attributed to the high capability of the VAE model to extract more relevant features compared to GAN, RBM, and DBN models by using variational inferences to approximate COVID-19 data probability distribution. Furthermore, it has been shown that the VAE combined with 1SVM has demonstrated the high capability for COVID-19 detection based on blood test data. The superiority of the VAE-1SVM approach could be associated with the flexibility and modeling capacity of VAE, from one side, and with the effectiveness of the 1SVM machine learning model. Importantly, in the VAE-1SVM, by mapping the extracted features obtained from the VAE model in a higher space, the projected features become linearly separable, and the detection problem becomes easy. In addition, it should be noted that addressing the COVID-19 infection detection as anomaly detection has improved the obtained results by focusing on only noninfected (normal) cases to effectively distinguish the abnormal (infected) cases.

C. Comparison With the State of the Art

Now, we compare the achieved performance of the developed VAE-Softmax and VAE-1SVM detectors with state-of-the-art methods applied to Datasets 1 and 2 (Table VIII). The study in [21] applied machine learning methods, including RF, artificial neural network (ANN), LR, and Lasso-elastic-net regularized generalized linear (GLMNET) models to predict SARS-CoV-2 infection. ANN has achieved the best classification results with an AUC of 0.95. In [52], five machine learning models have been investigated for identifying the risk of positive COVID-19: NN, RF, gradient boosting trees (GBTs), LR, and SVM. The SVM showed the best accuracy with an AUC of 0.85; de Freitas Barbosa *et al.* [17] tested multilayer perceptron (MLP), SVM, RF, random tree (RT), Bayesian network (BN), and NB for aiding to support the diagnosis of COVID-19 based on blood tests. Unexpectedly, BN achieved the highest overall accuracy of 95.159%. From Table VIII, we note that all the above-mentioned state-of-the-art methods are supervised machine learning methods, which needs labeling data to perform classification. On the other hand, the proposed VAE-1SVM detector is an unsupervised

TABLE VIII
COMPARISON WITH THE STATE-OF-THE-ART METHODS BASED ON DATASETS 1 AND 2

Refs	Dataset	Model	Metrics
[21]	Dataset 1	RF, LR, GLMNET, and ANN	AUC=95
[52]	Dataset 1	NN, RF, GBT, LR, and SVM	AUC=85
[17]	Dataset 1	MLP, SVM, RT, RF, BN, and NB	Acc=95.15% Sens=96.8%, Spec=93.6%
[12]	Dataset 1	XGBoost	AUC=99.38
Ours	Dataset 1	**VAE-Softmax**	**AUC=99.85**
Ours	Dataset 1	**GAN-Softmax**	**AUC=99.53**
Ours	Dataset 1	**VAE-1SVM**	**AUC=99.6**
[15]	Dataset 2	DT-XGBoost	AUC=85
Ours	Dataset 2	**VAE-1SVM**	**AUC=99.3**

and deep learning method, allowing it to learn deeply relevant information in blood test data. In addition, Table VIII reveals that the VAE-based methods outperformed the state-of-the-art methods by achieving a satisfying detection performance for the two studied datasets.

V. CONCLUSION

Accurate detection of infected COVID-19 patients is a key enabler for timely intervention and for mitigating the pandemic transmission. Recently, routine blood tests have been widely used to initial screen COVID-19 patients. This is motivated by the relatively fast availability of blood tests in all patient care locations with less expensive. This study attempted to detect COVID-19 using a novel unsupervised hybrid deep learning method based on routine blood tests. Notably, we presented the COVID-19 detection as an anomaly detection problem without using labeled data (i.e., fully unsupervised). The proposed data-driven detector combines the flexibility and accuracy of the VAE features extractor and the extended capability of 1SVM in anomaly detection. Crucially, we applied the 1SVM detector to features extracted by VAE for detecting COVID-19 cases. We assessed the proposed VAE-based 1SVM detector using two sets of routine blood tests samples: from the AEH, São Paulo, Brazil, and the SRH, Milan, Italy. We gave comparisons of the designed detector with seven hybrid data-driven methods: GAN-, DBN-, and RBM-based 1SVM, and VAE, GAN, DBN, and RBM with softmax layer as a discriminator layer as well as with the standalone 1SVM. Results based on the two considered datasets showed that unsupervised deep learning-based models provided satisfactory detection performance. Furthermore, results reveal that the proposed approach delivered the highest accuracy compared to the other investigated models. Overall, results demonstrated that unsupervised deep learning-based 1SVM detectors could effectively identify patients with COVID-19 based on routine blood tests data. Therefore, this study offers a promising alternative to a more widely available identification of infected patients with COVID-19.

In this study, applying unsupervised deep learning to routine blood tests data was revealed to be a potential tool for COVID-19 detection, which could be employed for supporting clinical decisions. Note that the data used in this work are relatively small, and it is normalized via z-normalization, making access to values of the original features impossible. In future work, we plan to verify the feasibility of the proposed methods to larger sized data when it is available. Another important direction of improvement is to fuse information from different sources from patient history, including chest X-rays, clinical signs, and symptoms. Thus, incorporating this relevant information, if available, in the construction of the proposed deep learning-driven detectors could further enhance their detection accuracy.

REFERENCES

[1] M. Day, "COVID-19: Identifying and isolating asymptomatic people helped eliminate virus in Italian village," *BMJ*, vol. 368, Mar. 2020, Art. no. m1165.

[2] B. Wang, Y. Zhao, and C. L. P. Chen, "Hybrid transfer learning and broad learning system for wearing mask detection in the COVID-19 era," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–12, 2021.

[3] R. R. Sharma, M. Kumar, S. Maheshwari, and K. P. Ray, "EVDHM-ARIMA-based time series forecasting model and its application for COVID-19 cases," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–10, 2021.

[4] W. Wu, J. Shi, H. Yu, W. Wu, and V. Vardhanabhuti, "Tensor gradient Lo-norm minimization-based low-dose CT and its application to COVID-19," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–12, 2021.

[5] M. Salath *et al.*, "COVID-19 epidemic in switzerland: On the importance of testing, contact tracing and isolation," *Swiss Med. Weekly*, vol. 150, Mar. 2020, Art. no. w202205.

[6] M. J. Loeffelholz and Y.-W. Tang, "Laboratory diagnosis of emerging human coronavirus infections—The state of the art," *Emerg. Microbes Infections*, vol. 9, no. 1, pp. 747–756, Jan. 2020.

[7] V. M. Corman *et al.*, "Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR," *Eurosurveillance*, vol. 25, no. 3, Jan. 2020, Art. no. 2000045.

[8] D. Li *et al.*, "False-negative results of real-time reverse-transcriptase polymerase chain reaction for severe acute respiratory syndrome coronavirus 2: Role of deep-learning-based CT diagnosis and insights from two cases," *Korean J. Radiol.*, vol. 21, no. 4, pp. 505–508, Apr. 2020.

[9] T. Ai *et al.*, "Correlation of chest CT and RT-PCR testing for coronavirus disease 2019 (COVID-19) in China: A report of 1014 cases," *Radiology*, vol. 296, no. 2, pp. E32–E40, Aug. 2020.

[10] Y. Yang *et al.*, "Evaluating the accuracy of different respiratory specimens in the laboratory diagnosis and monitoring the viral shedding of 2019-ncov infections," *MedRxiv*, pp. 1–17, 2020. [Online]. Available: https://www.medrxiv.org/content/early/2020/02/17/2020.02.11.20021493, doi: 10.1101/2020.02.11.20021493.

[11] M. A. Alves *et al.*, "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.

[12] 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, Jan. 2020, Art. no. 100449.

[13] J. Wu *et al.*, "Rapid and accurate identification of COVID-19 infection through machine learning based on clinical available blood test results," *medRxiv*, 2020. [Online]. Available: https://www.medrxiv.org/content/early/2020/04/06/2020.04.02.20051136, doi: 10.1101/2020.04.02.20051136.

[14] T. Roland *et al.*, "Machine learning based COVID-19 diagnosis from blood tests with robustness to domain shifts," *medRxiv*, 2021. [Online]. Available: https://www.medrxiv.org/content/early/2021/04/09/2021.04.06.21254997, doi: 10.1101/2021.04.06.21254997.

[15] F. Cabitza *et al.*, "Development, evaluation, and validation of machine learning models for COVID-19 detection based on routine blood tests," *Clin. Chem. Lab. Med. (CCLM)*, vol. 59, no. 2, pp. 421–431, Feb. 2021.

[16] 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, pp. 1–12, Aug. 2020.

[17] V. A. de Freitas Barbosa *et al.*, "Heg. IA: An intelligent system to support diagnosis of COVID-19 based on blood tests," *Res. Biomed. Eng.*, pp. 1–18, Jan. 2021, doi: 10.1007/s42600-020-00112-5.

[18] H. S. Yang *et al.*, "Routine laboratory blood tests predict SARS-CoV-2 infection using machine learning," *Clin. Chem.*, vol. 66, no. 11, pp. 1396–1404, Nov. 2020.

[19] M. Kukar *et al.*, "COVID-19 diagnosis by routine blood tests using machine learning," *Sci. Rep.*, vol. 11, no. 1, pp. 1–9, Dec. 2021.

[20] S. Aktar *et al.*, "Machine learning approach to predicting COVID-19 disease severity based on clinical blood test data: Statistical analysis and model development," *JMIR Med. Informat.*, vol. 9, no. 4, Apr. 2021, Art. no. e25884.

[21] A. Banerjee *et al.*, "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.
[22] R. P. Joshi *et al.*, "A predictive tool for identification of SARS-CoV-2 PCR-negative emergency department patients using routine test results," *J. Clin. Virol.*, vol. 129, Aug. 2020, Art. no. 104502.
[23] P. Schwab, A. D. Schütte, B. Dietz, and S. Bauer, "Clinical predictive models for COVID-19: Systematic study," *J. Med. Internet Res.*, vol. 22, no. 10, Oct. 2020, Art. no. e21439.
[24] D. Ai, Y. Wang, X. Li, and H. Pan, "Colorectal cancer prediction based on weighted gene co-expression network analysis and variational auto-encoder," *Biomolecules*, vol. 10, no. 9, p. 1207, Aug. 2020.
[25] K. Feng, H. Qin, S. Wu, W. Pan, and G. Liu, "A sleep apnea detection method based on unsupervised feature learning and single-lead electrocardiogram," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–12, 2021.
[26] E. Data4u. (Mar. 28, 2021)*Diagnosis of COVID-19 and its Clinical Spectrum AI and Data Science Supporting Clinical Decisions*. Kaggle. Accessed: Jul. 24, 2021. [Online]. Available: https://www.kaggle.com/einsteindata4u/covid19
[27] M. Kermali, R. K. Khalsa, K. Pillai, Z. Ismail, and A. Harky, "The role of biomarkers in diagnosis of COVID-19—A systematic review," *Life Sci.*, vol. 254, Aug. 2020, Art. no. 117788.
[28] T. A. Khartabil, H. Russcher, A. van der Ven, and Y. B. de Rijke, "A summary of the diagnostic and prognostic value of hemocytometry markers in COVID-19 patients," *Crit. Rev. Clin. Lab. Sci.*, vol. 57, no. 6, pp. 415–431, Aug. 2020.
[29] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," *Stat*, vol. 1050, p. 1, Dec. 2014.
[30] A. Dairi, F. Harrou, Y. Sun, and S. Khadraoui, "Short-term forecasting of photovoltaic solar power production using variational auto-encoder driven deep learning approach," *Appl. Sci.*, vol. 10, no. 23, p. 8400, Nov. 2020.
[31] F. Harrou, A. Dairi, F. Kadri, and Y. Sun, "Forecasting emergency department overcrowding: A deep learning framework," *Chaos, Solitons Fractals*, vol. 139, Oct. 2020, Art. no. 110247.
[32] A. Zeroual, F. Harrou, A. Dairi, and Y. Sun, "Deep learning methods for forecasting COVID-19 time-series data: A comparative study," *Chaos, Solitons Fractals*, vol. 140, Nov. 2020, Art. no. 110121.
[33] G. Boquet, A. Morell, J. Serrano, and J. L. Vicario, "A variational autoencoder solution for road traffic forecasting systems: Missing data imputation, dimension reduction, model selection and anomaly detection," *Transp. Res. C, Emerg. Technol.*, vol. 115, Jun. 2020, Art. no. 102622.
[34] Y. Zerrouki, F. Harrou, N. Zerrouki, A. Dairi, and Y. Sun, "Desertification detection using an improved variational autoencoder-based approach through ETM-landsat satellite data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 202–213, 2021, doi: 10.1109/JSTARS.2020.3042760.
[35] Y. Kawachi, Y. Koizumi, and N. Harada, "Complementary set variational autoencoder for supervised anomaly detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 2366–2370.
[36] V. Alves de Oliveira *et al.*, "Reduced-complexity end-to-end variational autoencoder for on board satellite image compression," *Remote Sens.*, vol. 13, no. 3, p. 447, Jan. 2021.
[37] M. Rolinek, D. Zietlow, and G. Martius, "Variational autoencoders pursue PCA directions (by accident)," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12406–12415.
[38] C. Doersch, "Tutorial on variational autoencoders," 2016, *arXiv:1606.05908*.
[39] T. Chen, X. Liu, B. Xia, W. Wang, and Y. Lai, "Unsupervised anomaly detection of industrial robots using sliding-window convolutional variational autoencoder," *IEEE Access*, vol. 8, pp. 47072–47081, 2020.
[40] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *J. Amer. Stat. Assoc.*, vol. 112, no. 518, pp. 859–877, 2017.
[41] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*.
[42] F. Harrou, Y. Sun, A. S. Hering, M. Madakyaru, and A. Dairi, "Unsupervised deep learning-based process monitoring methods," in *Statistical Process Monitoring Using Advanced Data-Driven and Deep Learning Approaches*. Amsterdam, The Netherlands: Elsevier, 2021, pp. 193–223.
[43] B. Schölkopf, J. C. Platt, J. C. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Comput.*, vol. 13, no. 7, pp. 1443–1471, 2001.
[44] F. Harrou, A. Dairi, B. Taghezouit, and Y. Sun, "An unsupervised monitoring procedure for detecting anomalies in photovoltaic systems using a one-class support vector machine," *Sol. Energy*, vol. 179, pp. 48–58, Feb. 2019.
[45] A. Dairi, F. Harrou, M. Senouci, and Y. Sun, "Unsupervised obstacle detection in driving environments using deep-learning-based stereovision," *Robot. Auto. Syst.*, vol. 100, pp. 287–301, Feb. 2018.
[46] T. Cheng, A. Dairi, F. Harrou, Y. Sun, and T. Leiknes, "Monitoring influent conditions of wastewater treatment plants by nonlinear data-based techniques," *IEEE Access*, vol. 7, pp. 108827–108837, 2019.
[47] F. Harrou, A. Dairi, Y. Sun, and M. Senouci, "Statistical monitoring of a wastewater treatment plant: A case study," *J. Environ. Manage.*, vol. 223, pp. 807–814, Oct. 2018.
[48] A. Dairi, T. Cheng, F. Harrou, Y. Sun, and T. Leiknes, "Deep learning approach for sustainable WWTP operation: A case study on data-driven influent conditions monitoring," *Sustain. Cities Soc.*, vol. 50, Oct. 2019, Art. no. 101670.
[49] F. Harrou, A. Dairi, Y. Sun, and F. Kadri, "Detecting abnormal ozone measurements with a deep learning-based strategy," *IEEE Sensors J.*, vol. 18, no. 17, pp. 7222–7232, Sep. 2018.
[50] R. J. Little and D. B. Rubin, *Statistical Analysis With Missing Data*, vol. 793. Hoboken, NJ, USA: Wiley, 2019.
[51] S. Van Buuren and K. Groothuis-Oudshoorn, "Mice: Multivariate imputation by chained equations in R," *J. Stat. Softw.*, vol. 45, no. 3, pp. 1–67, 2011.
[52] A. F. de Moraes Batista, J. L. Miraglia, T. H. R. Donato, and A. D. P. C. Filho, "COVID-19 diagnosis prediction in emergency care patients: A machine learning approach," *MedRxiv*, to be published.
[53] V. A. de Freitas Barbosa *et al.*, "Heg.IA: An intelligent system to support diagnosis of Covid-19 based on blood tests," *Res. Biomed. Eng.*, pp. 1–18, 2021.

Abdelkader Dairi received the Engineering degree in computer science from the University of Oran 1 Ahmed Ben Bella, Oran, Algeria, in 2003, the Magister degree in computer science from the National Polytechnic School of Oran, Es Senia, Algeria, in 2005, and the Ph.D. degree in computer science from Ben Bella Oran1 University, Oran, in 2018.

From 2007 to 2013, he was a Senior Oracle Database Administrator (DBA) and an Enterprise Resource Planning (ERP) Manager. He is currently an Assistant Professor with the Department of Computer Science, University of Sciences and Technology of Oran–Mohamed Boudiaf, Bir El Djir, Algeria. He has over 20 years of programming experience in different languages and environments. His research interests include programming languages, artificial intelligence, computer vision, machine learning, and deep learning.

Fouzi Harrou (Member, IEEE) received the M.Sc. degree in telecommunications and networking from the University of Paris VI, Paris, France, in 2006, and the Ph.D. degree in systems optimization and security from the University of Technology of Troyes (UTT), Troyes, France, in 2010.

He was an Assistant Professor with UTT for one year and the Institute of Automotive and Transport Engineering, Nevers, France, for one year. He was a Post-Doctoral Research Associate with the Systems Modeling and Dependability Laboratory, UTT, for one year. He was a Research Scientist with the Chemical Engineering Department, Texas A&M University at Qatar, Doha, Qatar, for three years. He is currently a Research Scientist with the Division of Computer, Electrical and Mathematical Sciences and Engineering, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia. He is the coauthor of the book *Statistical Process Monitoring Using Advanced Data-Driven and Deep Learning Approaches: Theory and Practical Applications* (Elsevier, 2020). His current research interests include statistical decision theory and its applications, fault detection and diagnosis, and deep learning.

Ying Sun received the Ph.D. degree in statistics from Texas A&M University, College Station, TX, USA, in 2011.

She held a two-year post-doctoral research position at the Statistical and Applied Mathematical Sciences Institute and the University of Chicago, Chicago, IL, USA. She was an Assistant Professor with The Ohio State University Columbus, OH, USA, for a year before joining King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia, in 2014. At KAUST, she established and leads the Environmental Statistics Research Group, which works on developing statistical models and methods for complex data to address important environmental problems. She has made original contributions to environmental statistics, in particular in the areas of spatiotemporal statistics, functional data analysis, visualization, and computational statistics, with an exceptionally broad array of applications.

Dr. Sun received two prestigious awards: the Early Investigator Award in Environmental Statistics from the American Statistical Association and the Abdel El-Shaarawi Young Research Award from The International Environmetrics Society.