Heliyon 10 (2024) e35677

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Distribution-based detection of radiographic changes in pneumonia patterns: A COVID-19 case study

Sofia C. Pereira [a,b,*], Joana Rocha [a,b], Aurélio Campilho [a,b], Ana Maria Mendonça [a,b]

[a] Institute for Systems and Computer Engineering, Technology and Science (INESC-TEC), Portugal
[b] Faculty of Engineering of the University of Porto, Portugal

ARTICLE INFO

Keywords:
Coronavirus
Deep learning
Autoencoder
Anomaly detection
Distribution shift
X-ray

ABSTRACT

Although the classification of chest radiographs has long been an extensively researched topic, interest increased significantly with the onset of the COVID-19 pandemic. Existing results are promising; however, the radiological similarities between COVID-19 and other types of respiratory diseases limit the success of conventional image classification approaches that focus on single instances. This study proposes a novel perspective that conceptualizes COVID-19 pneumonia as a deviation from a normative distribution of typical pneumonia patterns. Using a population-based approach, our approach utilizes distributional anomaly detection. This method diverges from traditional instance-wise approaches by focusing on sets of scans instead of individual images. Using an autoencoder to extract feature representations, we present instance-based and distribution-based assessments of the separability between COVID-positive and COVID-negative pneumonia radiographs. The results demonstrate that the proposed distribution-based methodology outperforms conventional instance-based techniques in identifying radiographic changes associated with COVID-positive cases. This underscores its potential as an early warning system capable of detecting significant distributional shifts in radiographic data. By continuously monitoring these changes, this approach offers a mechanism for early identification of emerging health trends, potentially signaling the onset of new pandemics and enabling prompt public health responses.

1. Introduction

Since 2020, numerous studies have focused on the automated classification of COVID-19 cases from Chest X-ray Radiographs (CXRs) [3,4,35], yielding promising outcomes. Later, it became clear that there are several limitations to distinguishing CXRs of COVID-19 patients from those with other types of pneumonia [30]. Issues such as shortcut learning and dataset shifts, often due to each data class being collected from different sources, were key factors behind apparent successes in COVID-19 classification algorithms [11,32]. Traditionally, medical diagnostic systems have relied on binary or multi-class classification schemes to differentiate between classes. One-class learning diverges from this by offering a unique framework to model the normal variations within a dataset, enabling the detection of any deviant pattern that signals change from the known distribution. This is especially useful for identifying emergent diseases, distinct from those previously known, such as what happened when the COVID-19 disease emerged.

* Corresponding author at: Institute for Systems and Computer Engineering, Technology and Science (INESC-TEC), Portugal.
E-mail address: sofia.c.pereira@inesctec.pt (S. C. Pereira).

https://doi.org/10.1016/j.heliyon.2024.e35677
Received 15 May 2024; Received in revised form 28 June 2024; Accepted 1 August 2024

Available online 5 August 2024

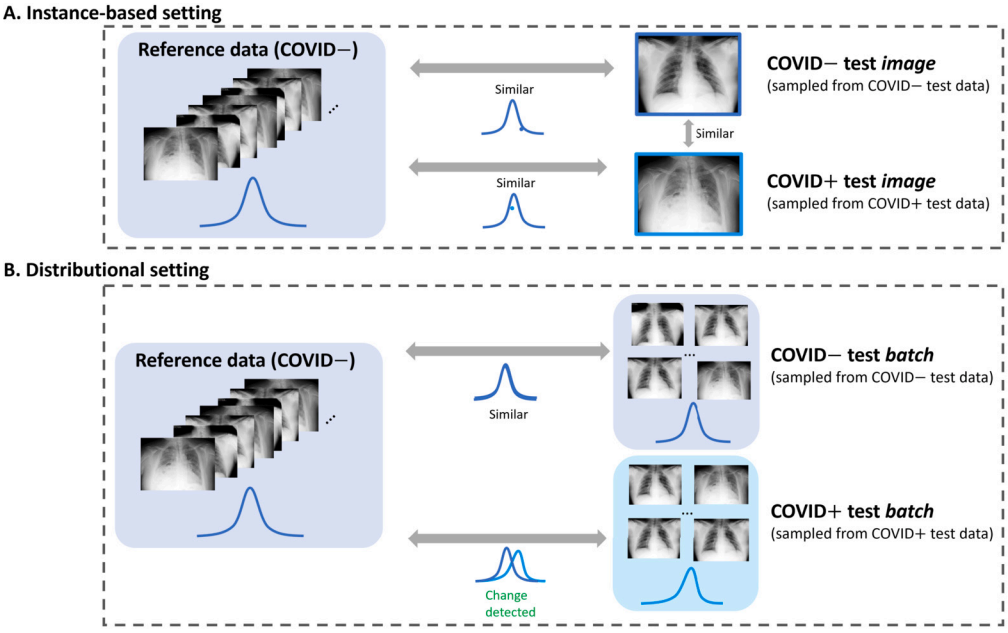


Fig. 1. Comparison of instance-based and distributional methods for detecting radiographic changes associated with COVID-19 pneumonia. Panel A illustrates the overlap of individual COVID-negative and positive radiographs with the reference set of images (COVID-negative), highlighting the challenge of distinguishing cases on a per-image basis (the dot overlay on the distribution curve represents the position of a single test image against the reference distribution). Panel B showcases how distributional analysis can differentiate batches of COVID-negative and positive radiographs, demonstrating its superior capability in discerning between typical known (similar to reference data) and unknown (different from reference data) radiographic patterns.

The radiological findings in CXRs of COVID-19 patients can be similar to those seen in patients with other kinds of pneumonia, especially those caused by other viruses [9,38]. The CXRs of COVID-19 patients often show characteristic signature patterns that are more prevalent within COVID-19 cases, such as a diffuse, peripheral, and commonly bilateral distribution of opacities, mostly in the lower region of the lungs [25]. However, these patterns may also be present in scans depicting other types of pneumonia. Therefore, classifying individual instances of COVID-19 pneumonia in CXRs is challenging because of a lack of discernible diagnostic features.

1.1. Overview

We propose a distributional OOD (Out-Of-Distribution) detection approach that evaluates sets of new instances collectively rather than individually. This method conceptualizes COVID-19 pneumonia as a distribution shift from traditional pneumonia, rather than as a completely new OOD class, thus accounting for the absence of exclusive diagnostic features specific to COVID-19 pneumonia. While it may be challenging to differentiate a single COVID-19 pneumonia CXR from conventional pneumonia cases, examining a population of COVID-19 pneumonia CXR scans allows us to verify that these cases are sampled from a new, shifted distribution. This strategy is visually represented in Fig. 1, which compares instance-based and distributional settings. Panel A illustrates the overlap of individual COVID-negative and positive radiographs with reference images, highlighting the difficulty of discerning cases on an individual basis. Conversely, Panel B highlights the effectiveness of the distributional method in differentiating between typical (similar to reference data) and shifted distributions (indicative of COVID-19), emphasizing its superior capability in detecting significant radiographic shifts. Our method is designed not to replace existing traditional or AI-based diagnostic methods that analyze images individually, but to *complement* them by offering a practical application in monitoring population-level shifts in radiographic data. This approach serves as a strategic tool for early detection of emerging trends and abnormalities, enabling timely interventions and informed public health measures. It provides actionable insights that aid in identifying significant changes that may warrant further investigation or public health measures.

1.2. Contributions

In this study, an autoencoder is used to extract compressed feature representations of each scan. We start by comparing scans with signs of pneumonia of suspected COVID-19 cases, categorized as COVID-positive (COVID+) and COVID-negative (COVID−), using instance-based approaches widely adopted in existing literature, such as examining the autoencoder's reconstruction loss and calculating the Mahalanobis distance in the feature space. We then take a novel distributional perspective, where the Maximum Mean Discrepancy (MMD), a well-known metric to compare data distributions, is computed between sets of COVID+ and COVID− pneumonia scans. Finally, we illustrate how the proposed methodology can serve as a warning mechanism for detecting changes in the radiographic manifestation of pneumonia. Our contributions can be summarized as follows:

- We demonstrate that COVID-19 CXRs are not OOD in an instance-based manner when compared with CXRs depicting non-COVID-19 pneumonia, as traditional instance-based approaches show poor performance.
- We show the superiority of a distributional methodology over an instance-based approach for effectively identifying COVID-19 CXRs, by introducing a novel framework that uses MMD to detect changes in populations of CXRs.
- We demonstrate the effectiveness of the proposed approach in identifying emerging COVID-19 cases and monitoring deviations in radiographic patterns by simulating an unsupervised drift detection framework within a data stream context.

2. Related work

Computer-Aided Diagnosis (CAD) systems, especially those utilizing Deep Learning (DL) methods like Convolutional Neural Networks (CNNs), have greatly advanced medical image analysis [41]. With the availability of large-scale annotated CXR datasets [18,39], there has been significant progress in developing DL models for disease classification [26,42]. The COVID-19 pandemic further accelerated research on automated lung CXR analysis [4,34]. However, some studies faced challenges such as dataset construction flaws and biases, limiting the discriminative performance of models [11,32,30].

Many studies adopted the conventional approach of using CNNs or transformers for individual image classification in COVID-19 detection, treating COVID-19 pneumonia cases as an independent data class distinct from other non-COVID-19 pneumonias [2,4]. Alternatively, this problem can also be framed as a one-class learning setting, where anomaly detection techniques such as those based on an autoencoder's reconstruction loss [1,21] or on the Mahalanobis distance [27,40] are employed to detect anomalous (or OOD) samples. Compared to traditional classification approaches, anomaly detection frameworks improve versatility by aiming to detect any kind of deviation from the data on which a model is trained. By training models on non-COVID-19 data only, the detection of any kind of change from the pneumonia patterns in which the model was trained can be assessed, and not just those specific to COVID-19 pneumonia.

In medical imaging, *dataset drift* is a common issue [14,15], occurring when real-world data statistical properties diverge from training data, leading to performance decline. Drift may arise from variations in imaging equipment, patient demographics, or imaging protocols, along with actual changes in disease manifestation. This underscores the necessity of continuous data and model monitoring, as well as the development of strategies to adapt to these dynamic environments. The work of Rabanse et al. [31] investigated various methods for detecting dataset shifts, categorizing them into feature-based, classifier-based, and density ratio techniques. They use MMD as a robust metric for detecting distributional shifts due to its effectiveness in comparing complex, high-dimensional, and noisy data distributions. This metric has been successfully applied in various domains, such as bioinformatics [43], finance [7] and medicine [12], demonstrating its robustness and flexibility in machine learning for distribution comparison tasks [13]. Taking some of these concepts and applying them to the medical domain, Soin et al. [33] underscore the importance of real-time drift detection in medical imaging, studying medical dataset shift in a data stream context. They aimed to define drift metrics that correlate with the deterioration of a classifier's performance, thereby serving as a warning mechanism for model deterioration. Koch et al. [23] demonstrated how traditional methods for detecting OOD data points fail to capture subgroup shifts in histopathology images, arguing that such shifts can be instead detected on a population level by treating them as a distribution shift.

Building on the concepts introduced in this section, our work frames COVID-19 pneumonia as a distribution shift compared to typical pneumonia cases. Unlike existing methods that focus on classifying individual COVID-19 pneumonia images, whether in a traditional classification approach or an anomaly detection setting, the novelty of our work lies in the application of MMD to detect distribution shifts in a population of pneumonia chest radiographs, using COVID-negative pneumonia scans as a reference distribution. We demonstrate that our method can detect significant shifts towards a new distribution of COVID-positive scans by analyzing sets of images rather than individual ones. This population-based approach enables the early detection of emerging patterns that may indicate public health concerns or outbreaks.

Our approach differs from existing literature in two key aspects: firstly, by conceptualizing COVID-19 pneumonia as a distribution shift relative to non-COVID-19 pneumonia, rather than as a completely distinct data class; and secondly, by focusing on detecting COVID-19 at the population level instead of the individual instance level. We assess this shift using feature-based methods, both in scenarios with complete data availability, similarly to those in [31], and in a streaming context, where data is continuously analyzed as it becomes available, similarly to [33]. Contrary to the model-centric approach of [33], where metrics correlating with classifier performance were proposed, our vision is data-centric, focusing on detecting drift in the data itself, independently of any classifier. In summary, our goal is to identify data drift, enabling the identification of population-level changes, while staying agnostic to model performance.

3. Data

We used two datasets: the BIMCV-COVID19 [16,17], employed in the majority of the experiments, and the BIMCV-COVID19-PADCHEST dataset [5] to perform experiments with external data. We chose this external data set because it was curated by the same authors and belongs to the same hospital network as BIMCV-COVID19, and both datasets were labeled using the same Unified Medical Language System (UMLS) tags [5], thus minimizing dataset shifts due to different image sources or inconsistent labeling. The UMLS tags were derived from radiology reports and clinical notes through a semi-automated process [16,17]. While both datasets belong to the Image Bank of the Valencian Community (BIMCV), the first consists of images from multiple hospitals, while the latter only contains images from one hospital. Both datasets and their preprocessing are described below. Table 1 shows the number of patients, sessions (individual interactions between the patient and the healthcare provider), and scans included in the study after

Table 1
Datasets used in this study (after preprocessing and filtering). Note that all patients of BIMCV-COVID19-PADCHEST are COVID−, as this data set was collected prior to the pandemic.

Data set	Cohort	Patients	Sessions	Scans
BIMCV-COVID19	+	999	1,094	1,374
BIMCV-COVID19	−	1,344	1,636	2,072
BIMCV-COVID19-PADCHEST	−	6,543	8,652	13,410

data preprocessing and filtering. All images were resized such that the smaller edge is 256 pixels, preserving the aspect ratio, and then center-cropped to 224 × 224 pixels (a size that balances computational efficiency while preserving enough image details).

3.1. BIMCV-COVID19

This dataset contains computed tomography and X-ray images, accompanied by the corresponding metadata, from patients observed between April 1^{st} 2020 and June 30^{th} 2020, sourced from the BIMCV. These patients were under clinical observation due to suspected SARS-CoV-2 infection, for which a microbiological Polymerase Chain Reaction (PCR) or serological test was requested. Additionally, each patient underwent at least one frontal CXR within a week of a microbiological or serological test. The patients were categorized into two cohorts: COVID+, indicating those with at least one positive test result (PCR or serological), and COVID−, representing those with consistently negative results on all tests (PCR or serological). In addition to the UMLS tags, the "COVID-19" and "COVID-19 uncertain" (denoting high or low suspicion of COVID-19, respectively) labels were also present. Only the data relative to CXR images were selected. The images were provided in Portable Network Graphic (PNG) format. We discarded images if (1) pixel data were not readable, (2) metadata were missing, (3) from pediatric patients, or (4) they were corrupted, were of a body part other than the thorax, or were of insufficient quality (low signal-to-noise ratio). For the images included in the study, we performed the preprocessing steps outlined in Appendix A.

To ensure that both cohorts (COVID+ and COVID−) only contain scans that are relevant to the problem, the data was filtered. This step helps focus the dataset on relevant cases and ensures that all included images contain characteristic features pertaining to pneumonia. Both cohorts were filtered by the most frequent tags that relate to pneumonia: *Increased density, Pneumonia, Alveolar pattern, Interstitial pattern, Infiltrates, Consolidation, Ground glass*. The COVID+ cohort originally included images from months after a positive PCR that do not show signs of disease, and therefore we further restricted this cohort to images containing the "COVID-19" label and with an acquisition date that is up to ± 3 days from a positive PCR result. Additionally, we excluded images from the negative cohort containing the tags "COVID-19" or "COVID uncertain", as these may likely be false negative instances.

3.2. BIMCV-COVID19-PADCHEST

This dataset is a subset of the PadChest [5], curated by the authors of BIMCV-COVID19. The images were collected at San Juan de Alicante Hospital from 2009 to 2017 and were extracted from the BIMCV. All patients of BIMCV-COVID19-PADCHEST are COVID-negative, as this data set was collected before the pandemic. The authors curated this subset intending to select images containing findings related to COVID-19, to serve an additional COVID-negative cohort. The criteria used was the collection of CXRs tagged with *pneumonia, infiltrates*, or both. Images may be additionally tagged with other of the 174 labels in the PadChest dataset. A control group designated as *normal*, containing labels on other abnormalities unrelated to COVID-19 (such as *cardiomegaly, support devices*, or *fractures*) is also present. Corrupted and pediatric images were excluded and the data were filtered using the same tags used for filtering the COVID+ and COVID− cohorts of the BIMCV-COVID19 dataset, to ensure consistency.

4. Methods

In this section, we detail the methods, starting with the autoencoder used for obtaining the feature representations of data, followed by the notation and experimental setups used in the study.

4.1. Latent representation and reconstruction loss

To obtain a latent representation of each image, we use an autoencoder. This type of fully convolutional model relies on self-supervised learning, where representation learning is achieved through the task of reconstructing the network's input. An autoencoder comprises an encoder to transform input data into an efficient latent representation, and a decoder that aims to reconstruct the original input based on the bottleneck latent representation. After successful training, the model can generate meaningful latent representations of the inputs. We employ a pretrained autoencoder [10] that was trained on multiple CXR datasets [5,18,20,39] using an elastic loss (the sum of the mean absolute and mean squared errors). The images are rescaled to the $[-1024, 1024]$ range, which is a prerequisite of the `TorchXRayVision`, and then processed through the autoencoder to obtain their latent representation and reconstruction loss. The autoencoder's bottleneck layer has shape [512, 3, 3] and global average pooling is performed over the second and third dimensions to obtain a single 512-dimensional feature vector. These latent representations are used in all experiments detailed below.

Table 2
Approaches explored in this study and their main characteristics.

	Distributional Approach	Uses Feature Vector	Data are separated into reference and test sets	Assumes test data are unlabeled and become available through time
Reconstruction loss	✗	✗	✗	✗
Mahalanobis	✗	✓	✓	✗
MMD *bulk*	✓	✓	✓	✗
MMD *stream*	✓	✓	✓	✓

Table 3
Notation used throughout the manuscript.

Notation	Description
$D^+ = \{x_1^+, x_2^+, x_3^+, \dots, x_{n^+}^+\}$	Set of COVID+ scans from P^+ patients of BIMCV-COVID19
$D^- = \{x_1^-, x_2^-, x_3^-, \dots, x_{n^-}^-\}$	Set of COVID− scans from P^- patients of BIMCV-COVID19
$D^p = \{x_1^p, x_2^p, x_3^p, \dots, x_{n^p}^p\}$	Set of external COVID-19 negative scans from P^p patients of BIMCV-COVID19-PADCHEST
$P^+ / P^- / P^p$	Number of patients in $D^+ / D^- / D^p$
$set_r = \{x_1^{t_r}, x_2^{t_r}, x_3^{t_r}, \dots, x_{n_r}^{t_r}\}$	Reference set with scans of data type t_r
$set_{in} = \{x_1^{t_{in}}, x_2^{t_{in}}, x_3^{t_{in}}, \dots, x_{n_{in}}^{t_{in}}\}$	In-distribution test set with scans of data type t_{in}
$set_{out} = \{x_1^{t_{out}}, x_2^{t_{out}}, x_3^{t_{out}}, \dots, x_{n_{out}}^{t_{out}}\}$	Out-of-distribution test set with scans of data type t_{out}
$t_{in} / t_{in} / t_{in} \in \{+, -, p\}$	Data type in $set_{in} / set_{out} / set_p$
$b_{in} = \{x_1^{t_{in}}, x_2^{t_{in}}, x_3^{t_{in}}, \dots, x_B^{t_{in}}\}$	Batch of samples drawn from set_{in}
$b_{out} = \{x_1^{t_{out}}, x_2^{t_{out}}, x_3^{t_{out}}, \dots, x_B^{t_{out}}\}$	Batch of samples drawn from set_{out}
$R \in]0, 1[$	Reference percentage
$M \in \mathbb{N}_{>0}$	Number of b_{in} / b_{out} batches
$B \in \mathbb{N}_{>0}$	Batch size
$K \in \mathbb{N}_{>0}$	Number of Repetitions
$L \in \mathbb{N}_{>1}$	Window width
$S \in \mathbb{N}_{>1}$	Window stride
$S_0 \in [0, n_{out}]$	Number of infected patients in set_{out} at time step 0
$R_0 \in \mathbb{R}_{>0}$	Basic reproduction number
$F \in \mathbb{N}_{>0}$	Update frequency
$d_M(Q, \vec{x})$	Mahalanobis distance between distribution Q and point x
$MMD(P, Q)$	Maximum mean discrepancy between distributions P and Q

4.2. Problem setup and notation

Table 2 provides an overview of the approaches explored in this study, and Table 3 summarizes the notation used throughout the manuscript. As a baseline, we start by implementing conventional anomaly detection approaches previously described in the literature for this problem, namely, monitoring via reconstruction loss and Mahalanobis distance, applied to our curated version of BIMCV-COVID19. Subsequently, the proposed distributional approach is introduced, which relies on calculating the Maximum Mean Discrepancy (MMD) between two populations of data as detailed below.

For analyzing the reconstruction loss, the entire dataset is used collectively. For the remaining experiments, involving the Mahalanobis distance and MMD, the data are divided into reference data (set_r) and test data. Batches of test data of size B are compared to set_r. The *in-distribution* test set (set_{in}) generates M in-distribution test batches (b_{in}), while the OOD test set (set_{out}) generates M OOD test batches (b_{out}). The in-distribution test batches do not represent a drift from set_r, while the OOD test batches are assumed to contain drifted data. The specific data (COVID-positive or negative pneumonia) contained in these sets depends on the specific experiment, with the main experiment defining the reference data and in-distribution test data as COVID-negative pneumonia, and the OOD test data as COVID-positive pneumonia, allowing for the effective detection of distributional shifts. For the Mahalanobis distance experiments (instance-based approach), data batches contain a single image, while for the MMD experiments (distributional approach), batches contain sets of images and are drawn with replacement.

Each of the three components detailed above (set_r, set_{in}, and set_{out}) can comprise COVID-negative data (D^-, 2,072 scans of 1,344 patients), COVID-positive data (D^+, 1,374 scans of 999 patients), and alternatively, PadChest data (D^p, 13,410 scans of 6,543 patients, all COVID-19 negative) for experiments involving external data. The reference percentage $R \in [0, 1]$ determines the proportion of a given data type (D^-, D^+, or D^p) to be utilized for generating set_r. The remaining $1 - R$ patients from the same data type, as well as patients from the other data types, remain available to constitute both set_{in} and set_{out}. To eliminate unwanted sample size bias, set_{in} and set_{out} are made to have scans from the same number of patients. In the default scenario mentioned above, set_r is drawn from D^-, set_{in} comprises the scans of the remaining individuals from D^-, and set_{out} contains data from D^+. This configuration can be flexibly altered in multiple ways to observe the system's response and adaptations.

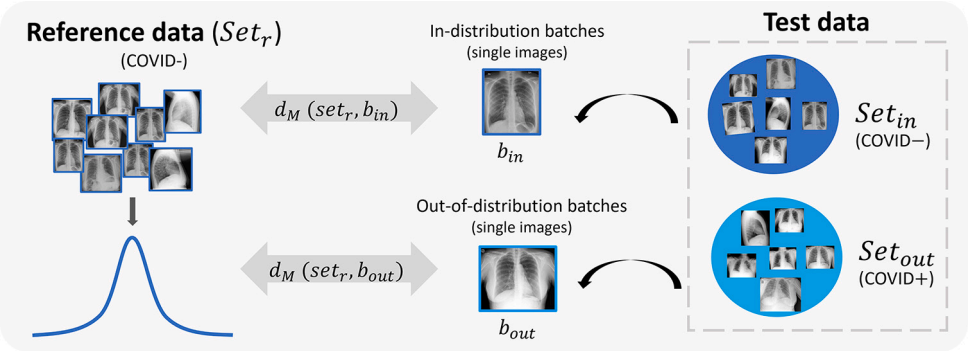


Fig. 2. Implementation of the Mahalanobis distance (d_M) calculation between reference data and individual test samples.

4.3. Instance-based experiments

We start by analyzing the reconstruction loss obtained from the autoencoder, a common method to detect OOD instances. A low reconstruction loss indicates similarity to the instances seen during training, while a large loss indicates difficulty in reconstructing the inputs differing from those seen in the training set. Assuming that COVID+ pneumonia instances are different from previous pneumonia cases in an instance-wise manner, we expect a higher reconstruction loss for CXRs depicting COVID-19 pneumonia compared to non-COVID-19 pneumonia cases, since the autoencoder's training set does not contain COVID+ cases. We pass all D^+ and D^- instances through the autoencoder and compare their reconstruction loss. The Area Under the receiver operating characteristic Curve (AUC) between the losses of D^- and D^+ scans is used for assessing the ability to distinguish scans. Although the AUC is typically associated with probability metrics, it can be applied to distance metrics to evaluate their ability to discriminate between different classes based on proximity. In this case, the AUC is calculated based on the true positive and false positive rates across different reconstruction loss thresholds. Ideally, the reconstruction loss of COVID+ cases would always exceed that of COVID− cases, resulting in an AUC of 1.0, while a random or poorly performing metric would yield an AUC of 0.5.

The last and main instance-based assessment relies on the Mahalanobis distance. This metric represents the distance between a data point and a data distribution in a multi-dimensional space. It measures how many standard deviations away a data point is from the mean of the distribution, adjusted for the correlations between the variables. Similarly to the reconstruction loss, it is also a popular choice of metric for detecting OOD samples, with the advantage that it leverages the vector latent representation of an image, instead of a single scalar. Given a probability distribution Q on $\mathbb{R}^N$, with mean $\vec{\mu} = (\mu_1, \mu_2, \mu_3, \ldots, \mu_N)^\top$ and positive-definite covariance matrix C, the Mahalanobis distance of a point $\vec{x} = (x_1, x_2, x_3, \ldots, x_N)^\top$ from Q is presented in equation (1).

$$d_M(Q, \vec{x}) = \sqrt{(\vec{x} - \vec{\mu})^\top C^{-1} (\vec{x} - \vec{\mu})} \tag{1}$$

For this approach, the data are divided into reference and test sets as detailed above, and each test batch is comprised of a single image ($B = 1$) without replacement. Under these conditions, the Mahalanobis distance [24] (Equation (1)) is calculated between each test batch (i.e., each image in the test data) and set_r. The implementation is detailed in Fig. 2, where $t_{ref} = -$, $t_{in} = -$, $t_{out} = +$ and $R = 0.5$. Since the data are limited, we perform the experiments K times, resulting in different data partitions being randomly selected for each set of data each time, making the results more robust. We then compute the AUC between the d_M of b_{in} and b_{out} batches.

4.4. Distributional experiments

In the proposed distributional setting, batches containing more than one image ($B > 1$) are drawn with replacement. Our approach and design were based on that in [37]. The MMD [13] serves as a kernel-based statistic for comparing two probability distributions $P(X)$, with samples $\{x_i\}_{i=1}^n$, and $Q(Y)$, with samples $\{y_j\}_{j=1}^m$). In our experiments, a Gaussian kernel $k(x, y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$ is used (Equation (2)). The parameter σ controls the spread of the Gaussian kernel, which was set using the median heuristic, following [13]. While Mahalanobis distance relies on the assumption of data following a Gaussian distribution, MMD does not impose assumptions on the underlying data distribution, which can be advantageous when dealing with complex and irregular distributions that are often encountered in medical imaging. This flexibility makes MMD a suitable choice for our context.

$$MMD(P, Q) = \frac{1}{n(n-1)} \sum_{i \neq j}^{n} k(x_i, x_j) + \frac{1}{m(m-1)} \sum_{i \neq j}^{m} k(y_i, y_j) - \frac{2}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} k(x_i, y_j) \tag{2}$$

The MMD between set_r and each b_{in} (*no drift*) and b_{out} (*drift*) batch is calculated. We implement the proposed distributional approach under two settings, which are further detailed below:

- **Bulk:** This setting considers that (1) all data are available at a given time point and (2) that we have *a priori* access to labeled drifted data (COVID-19 positive scans, in this case). The MMDs of all b_{in} and b_{out} batches are used to calculate the AUC.

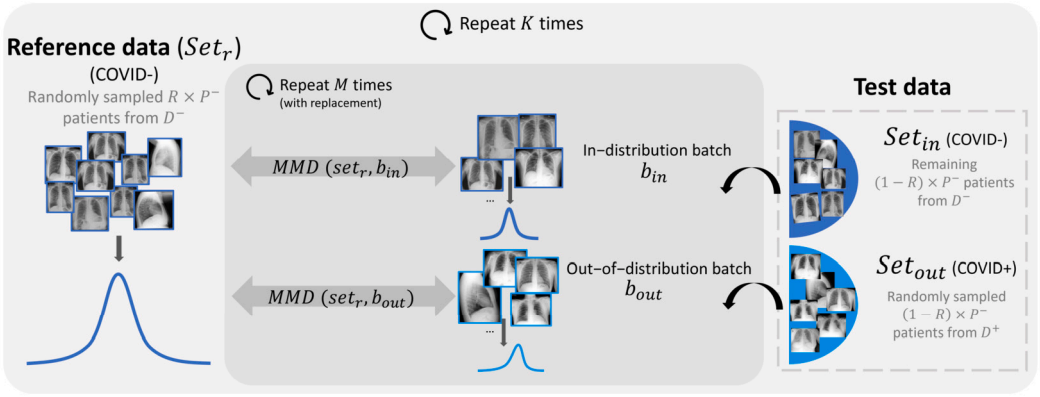


Fig. 3. Diagram of the Maximum Mean Discrepancy (MMD) calculation between reference data and batches of test data in *bulk*.

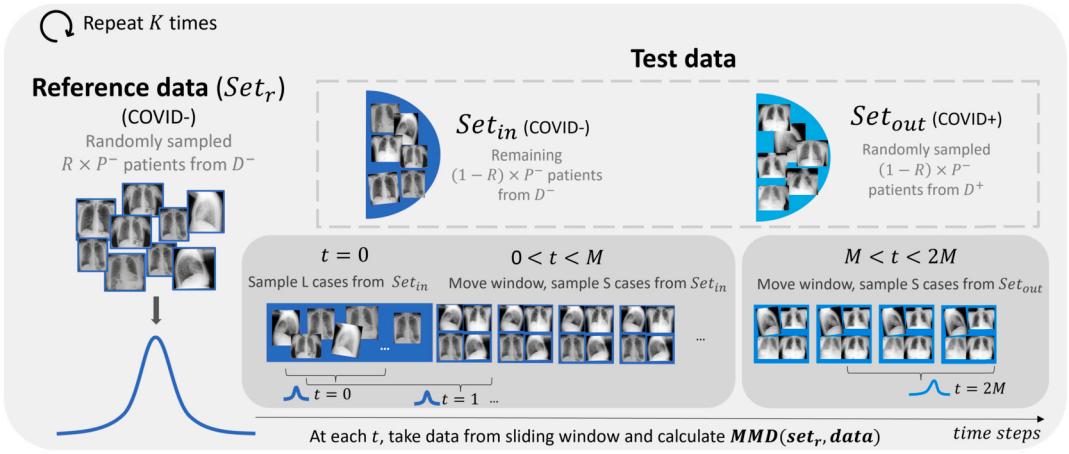


Fig. 4. Diagram of the Maximum Mean Discrepancy (MMD) calculation between reference data and batches of test data in *stream*.

- **Stream:** This scenario tries to better mimic a real epidemiological setting, where data are generated throughout time. This scenario considers that (1) we do not have access to all the data at a given time point, but rather that data become available as time passes and (2) we do not have *a priori* access to labeled drifted data. This implies that drift detection needs to be performed for a single data batch of unknown nature at each point in time. The MMD is used directly as a test statistic to perform statistical hypothesis testing for each batch, avoiding the need for labeled drifted data.

4.4.1. Bulk setting

In this approach, represented in Fig. 3, M b_{in} batches are drawn from set_{in} and M b_{out} batches are drawn from set_{out}. Then, the MMD between each test batch and set_r is calculated and, finally, the results are evaluated using the AUC between the MMD of b_{in} and b_{out} test batches. The default configuration in Fig. 3 has $t_r = -$, $t_{in} = -$, and $t_{out} = +$, being therefore expected that b_{in} batches are closer to the reference set than b_{out} batches. Additional experiments explore scenarios where the data types of the sets are changed. For example, a scenario where $t_r = -$, $t_{in} = -$, and $t_{out} = -$ provides valuable insights into the concept of *baseline drift*, which refers to the anticipated level of separability between batches of the same data type, that may exhibit a non-zero offset due to inherent data characteristics. The experiments are implemented with parameters $K = 50$, $R = 0.5$, $M = 100$, and $B = 50$. The sensitivity analysis on these hyperparameters is shown in Section 5.3.

4.4.2. Stream setting

The *stream* approach only presupposes access to set_r at the initial time step (0), simulating a data stream where additional data are introduced over time, according to a window size L and stride S (Fig. 4). Multiple test batches form a detection window, that moves in a sliding fashion. At each time step $t = (t_1, t_2, t_3, \ldots, t_{2M})$, a new data batch is pooled. At $t = 0$, L new samples form the initial detection window. For $t > 0$, S new samples are pooled, discarding the first S samples from the previous window, creating a new detection window. In the first M time steps, data batches b_{in} are pooled from set_{in}, while during the last M time steps, data batches b_{out} are pooled from set_{out}, to simulate the emergence of a COVID-19-induced drift.

At each time step, we perform statistical hypothesis testing between the current detection window and the reference set, producing a new MMD value corresponding MMDs at the confidence levels $\alpha = \{0.01, 0.05, 0.1\}$. This is accomplished through a permutation

test, where under the null hypothesis, $P = Q$. Intermingling data from both distributions to create two new distributions, P' and Q', should result in MMD values similar to that between data from P and Q. By reshuffling data samples and recalculating the MMD, the Cumulative Distribution Function (CDF) of the MMD under the null hypothesis is estimated. If the original MMD value exceeds the $1 - \alpha$ quantile of the empirical cumulative distribution function, we reject the null hypothesis; otherwise, we accept it. We perform 1000 permutations to draw the CDF under the null hypothesis. The experiments are implemented with parameters $L = 100$, $S = 25$, and $M = 50$, guided by the results of the sensitivity analysis performed in the *bulk* experiments, described in Section 5.3. We also explore varying S, to see its impact on the results (Fig. 8 in Section 5.3).

4.4.3. Drift purity

The initial experimental setup utilizes set_r and set_{in} containing only COVID− samples, with set_{out} comprising exclusively COVID+ samples. To replicate real-world scenarios, particularly at a pandemic's onset, we modify set_{out} to include COVID− scans, thereby reducing the purity of COVID+ cases in this set, which is essential for assessing detection capabilities under varying conditions. The COVID+ proportion in set_{out} is adjusted to values ranging from 95% to 50%, influencing the composition of b_{out} batches. This adjustment also affects the size of set_{in} due to the limited patient data available.

$$Every\ F\ time\ steps: \quad S_t = S_0[1 + R_0^t + t(R_0^{t-1} + \cdots + R_0^1)] \tag{3}$$

We explore both static and dynamic purity adjustments: static changes are implemented under the *drift* setting (section 4.4.1) and maintain a constant COVID+ ratio in set_{out}, whereas dynamic changes, implemented under the *stream* setting (section 4.4.2), exponentially increase COVID+ cases in set_{out} to simulate pandemic growth. To align our simulations with concepts from the field of public health, the exponential growth function is defined with three key parameters: the initial number of infected patients S_0, the basic reproduction number R_0 and the frequency of the updates F. The R_0 is the average number of new cases stemming from a single infected individual. In real-world situations, this value is influenced by multiple external factors, including patient isolation and vaccination. For the sake of simplicity, our simulation disregards these external influences. Initially, both set_{in} and set_{out} consist entirely of COVID− scans. As time steps progress, the number of COVID+ cases in set_{out} is increased according to Equation (3). Parameters S_0, R_0 and F are initially set to 1, 2 and 5, respectively. In comparison to the previous experiments, R was decreased from 0.5 to 0.3, guided by the sensitivity analysis in Fig. 7 (Section 5.3), to increase the availability of negative samples in the set_{out} as purity decreases.

4.4.4. External data

To assess the generalization of the findings to other datasets, an external data source, the BIMCV-COVID19-PadChest (D^p) dataset, was incorporated into the experimental framework. As detailed in Section 3, this dataset originates from a single hospital within the broader BIMCV network, whereas the BIMCV-COVID19 dataset comprises data collected from multiple hospitals within the network. Given that D^p is a dataset collected before the emergence of the COVID-19 pandemic, it is reasonable to foresee that the distribution of D^- data aligns more closely with the distribution of D^p than that of D^+. To test this hypothesis, external data are used to form set_r and set_{in} in some of the developed experiments under the *bulk* setting. In such cases, R was reduced to 0.1 to enhance computational efficiency and align the size of D^p more closely to D^+ and D^-. This external dataset is also introduced into the *stream* simulations, to build set_r, similarly as described above.

5. Results

In this section, we present and analyze the results of the experiments detailed in section 4, which aimed to identify the most effective method for distinguishing between scans of COVID+ and COVID− pneumonia.

5.1. Instance-based approaches

The first instance-based approach leverages the autoencoder's reconstruction loss as a surrogate metric for detecting OOD scans. Given that the autoencoder was trained on data that did not contain COVID+ cases, and under the assumption that individual COVID+ scans exhibit instance-wise distinctive characteristics compared to COVID−, a higher reconstruction loss for COVID+ scans would be expected. However, our analysis yielded results contrary to this expectation, as there are no relevant differences in reconstruction loss between COVID− and COVID+ samples, evidenced by an AUC close to 0.5 reported in Table 4.

Evaluating the separability of single instances using the derived feature space may offer more comprehensive insights compared to relying solely on the reconstruction loss, given that features encompass a multidimensional space (512 dimensions) rather than a single scalar. The AUC of the Mahalanobis distance (where $R = 0.5$, $K = 50$, $B = 50$, $M = 100$, $t_r = -$, $t_{in} = -$ and $t_{out} = +$) for b_{in} and b_{out} scans relative to set_r set is also presented in Table 4. The results obtained using the features are consistent with those obtained with the reconstruction loss. Collectively, these findings support our initial hypothesis that COVID+ scans do not exhibit significant instance-wise differences compared to COVID− scans.

5.2. Distributional approaches

The results of the distributional approach are presented in the following subsections. We begin with the *bulk* analysis results, followed by those of the *stream* analysis.

Table 4
AUC for the reconstruction loss and Mahalanobis distance metrics. Mean ± standard deviations are presented, where applicable.

Metric	AUC
Reconstruction loss	0.463
Mahalanobis distance[a]	0.469 ± 0.012

[a] $t_r = -$, $t_{in} = -$ and $t_{out} = +$.

Table 5
AUC obtained for the distributional *bulk* approach (Fig. 3). $K = 50$, $M = 100$, $B = 50$ and $R = 0.5$. Multiple combinations of t_r, t_{in} and t_{out} are tested. The default scenario is shown in bold.

t_{ref}	t_{in}	t_{out}	n_r	n_{in}	n_{out}	AUC
−	−	+	1,032 ± 16	1,040 ± 16	925 ± 12	**0.938 ± 0.031**
−	−	−	1,031 ± 17	520 ± 15	522 ± 16	0.497 ± 0.063
−	+	−	1,032 ± 16	925 ± 12	1,040 ± 16	0.064 ± 0.033
−	+	+	1,037 ± 17	687 ± 8	687 ± 8	0.488 ± 0.129

Table 6
AUC obtained for the distributional *bulk* approach, when external data is used for t_r and, optionally, t_{in}. Multiple combinations of t_r, t_{in} and t_{out} are tested. $K = 50$, $M = 100$, $B = 50$ and $R = 0.1$.

t_{ref}	t_{in}	t_{out}	n_r	n_{in}	n_{out}	AUC
p	p	−	1,340 ± 35	2,755 ± 46	2072 ± 0	1.000 ± 0.000
p	p	+	1,333 ± 25	2,040 ± 33	1,375 ± 0	1.000 ± 0.000
p	−	−	1,341 ± 31	1,034 ± 18	1,038 ± 18	0.509 ± 0.106
p	−	+	1,333 ± 25	1,542 ± 16	1,375 ± 0	0.978 ± 0.009
p	+	−	1,333 ± 25	1,375 ± 0	1,542 ± 16	0.024 ± 0.009
p	+	+	1,333 ± 31	686 ± 10	687 ± 10	0.500 ± 0.143

5.2.1. Bulk

Following the procedure outlined in section 4.4, we obtained the results presented in Table 5. In the most conventional scenario, characterized by $t_r = -$, $t_{in} = -$ and $t_{out} = +$, the MMD of the b_{in} batches is consistently smaller than that of the b_{out} batches, as underscored by the high AUC. Notably, when all sets of data are comprised of the − data type, the separability of MMD between b_{in} batches and b_{out} is close to 0.5, indicating no discriminant ability. Finally, when $t_r = -$, $t_{in} = +$ and $t_{out} = -$ or $t_r = +$, $t_{in} = -$ and $t_{out} = +$, the AUC is close to zero, as the MMD of the b_{in} batches is consistently larger than that of the b_{out} batches (the problem is "inverted"), again indicating a good result. We performed a sensitivity analysis on K, M, B and R (Section 5.3). The results of this analysis guided and supported the selection of hyperparameters.

Table 6 presents the results using external data D^p (BIMCV-COVID19-PadChest) as reference data. Some inherent and uncontrollable dissimilarities exist between BIMCV-COVID19-PadChest and BIMCV-COVID19, encompassing variations in data acquisition, scanning devices, etc., leading to dataset shift. Consequentially, the MMD between data from different sources will always be greater than that between data from the same source. As a consequence, when D^p is employed both for set_r and set_{in}, a very strong drift (AUC = 1.0) will consistently emerge relative to data from BIMCV-COVID19 in set_{out}, regardless of it being D^- or D^+ (first and second lines of Table 6), due to dataset shift. However, when D^p is used for set_r but set_{in} is made of COVID− data and set_{out} of COVID+, the AUC is still high, meaning that the MMD between set_r and set_{in} (with D^- data) batches is smaller than that between set_r and set_{out} (with D^+ data) batches, allowing their discrimination. This indicates that the distribution of D^- data is closer to D^p than to D^+. These results highlight the need for threshold calibration, further discussed in Section 6.

Considering the default scenario in Table 5, where $t_r = -$, $t_{in} = -$ and $t_{out} = +$, decreasing the drift purity (the percentage of COVID+ samples in set_{out}) implies that b_{out} batches will no longer consist solely of 100% COVID+ cases. Fig. 5 illustrates how reducing the purity of set_{out} leads to a decline in the AUC. This is expected because, as the drift's purity decreases, b_{out} batches become more similar to the COVID− reference set, leading to a reduced MMD. The MMD of the b_{in} and b_{out} relative to the reference set becomes more similar, thus reducing their separability. At the bottom of the figure, the sizes of set_r, set_{in} and set_{out} are displayed. While the size of set_r is constant, as defined by R, those of set_{in} and set_{out} fluctuate with changes in purity since lower purity requires the use of negative cases to construct the set_{out}, which, in turn, reduces the pool of negative samples available for creating the set_{in}. When purity is at 50%, set_{out} comprise $\frac{1}{2}$ COVID− and $\frac{1}{2}$ COVID+ cases and there is no discriminative power.

5.2.2. Stream

In all stream experiments, $t_{in} = -$ and $t_{out} = +$. Initially, we set $t_r = -$ and $R = 0.3$ (Fig. 6a), and subsequently switch to $t_r = p$ and $R = 0.1$ (Fig. 6b). In both cases, $L = 100$, $S = 25$ and $M = 50$. Section 5.3 explores the impact of varying S on the results. In addition

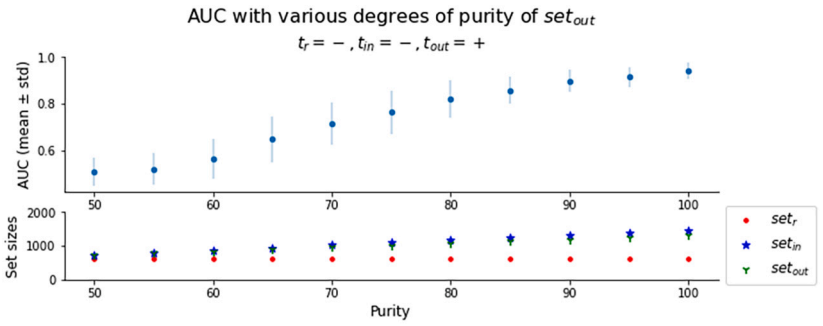
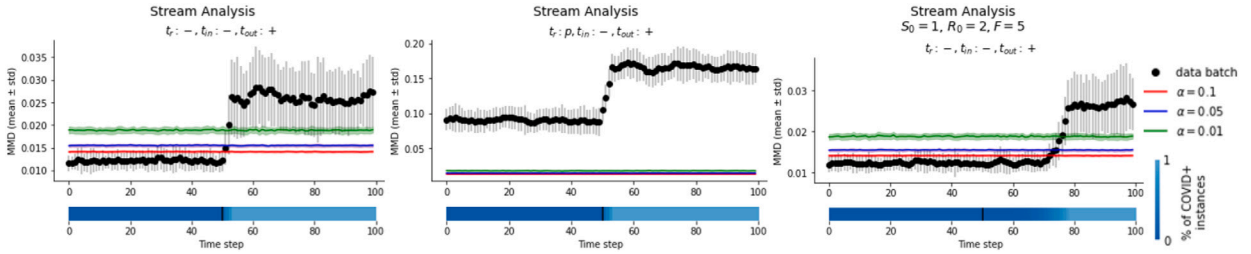


Fig. 5. Effect of changing drift purity in the AUC. The lower panel indicates the size of set_r ($R = 0.3$), set_{in} and set_{out} for different levels of purity.



(a) Standard setting (b) External data in set_r (c) Exponential growth of cases

Fig. 6. MMD of data batches over time steps ($L = 100$, $S = 25$ and $M = 50$). The lower bar shows the percentage of COVID+ patients in the data pool from which batches are being sampled (set_{in} during the first M steps and set_{out} thereafter).

to presenting the MMD score at each time step, the plots also display the MMD corresponding to the 0.1, 0.05, and 0.01 confidence thresholds at each time step. These values vary slightly over time because, as described in Section 4, they are calculated based on a permutation test between set_r and the data at time step t, which varies over time. In Fig. 6a, it is easy to discern the moment when the data transitions from being sampled from the set_{in} to being drawn from the set_{out}. Given that $S = 25$ and $L = 100$, this transition unfolds gradually over four time steps. The batches drawn from a distribution containing approximately 25% COVID+ scans (first transition point) exhibit MMD values below all the specified confidence thresholds. Conversely, for batches drawn from a distribution with approximately 75% COVID+ scans (last transition point), the MMD exceeds both the 0.1 and 0.05 confidence thresholds, which is consistent with Fig. 5, where a purity of 70% already results in an AUC value of approximately 0.7.

Unlike the previous AUC-based assessment from Section 5.2.1, here, MMD serves directly as a test statistic for hypothesis testing, making its magnitude relevant. When $t_r = p$, confidence thresholds, calculated under the null hypothesis, become inadequately low for deciding on test batches from another dataset due to dataset shift. In Fig. 6b, all confidence thresholds are surpassed. However, the MMD for COVID+ data is consistently larger than that for COVID− data. With proper calibration, it becomes feasible to establish an MMD threshold to distinguish b_{in} batches from b_{out} batches. This will be further addressed in Section 6.

Previously, in Section 5.2.1, the concept of *drift purity* was also explored. However, since the temporal progression of data was not a component in the *bulk* experiments, the purity of the set_{out} was adjusted in a static manner. Here, in the *stream* simulations, where the temporal progression of data is considered, dynamic purity changes were implemented. Fig. 6c illustrates how, compared to past experiments, the impact of drift becomes noticeable later, as purity is incrementally increased slower and over a longer period of time, following Equation (3), with $S_0 = 1$, $R_0 = 2$, and $F = 5$ (there is one initial case, and every five steps, each case generates an additional two cases). Other simulations with different values of I, R and F are in Section 5.3.

5.3. Sensitivity analyses

This section presents the results for the sensitivity analyses. For the *bulk* experiments, we analyze how the AUC varies when varying each hyperparameter in the following ranges: $K : \{5, 10, 25, 50, 75, 100\}$ (default: 50); $B : \{5, 10, 25, 50, 75, 100\}$ (default: 50); $M : \{1, 25, 50, 75, 100, 125, 150, 175, 200\}$ (default: 100); $R : \{0.1, 0.25, 0.5, 0.75, 0.9\}$ (default: 0.5). The analyses provide valuable insights on the effects of hyperparameters, summarized in Fig. 7. This figure shows that the patterns observed in Table 5 are robust across the range of hyperparameter values. Fig. 8 shows the effect of increasing (a) or decreasing (b) the stride on *stream* simulations, and Fig. 9 shows two additional configurations for the exponential growth of cases. Note that for Fig. 8b, M was increased to 100 to allow the last batches to achieve a purity of 100%. Note the relationship between L and S parameters in the *stream* experiments and B in the *bulk* experiments; the sensitivity analyses performed for B informed the choice of L.

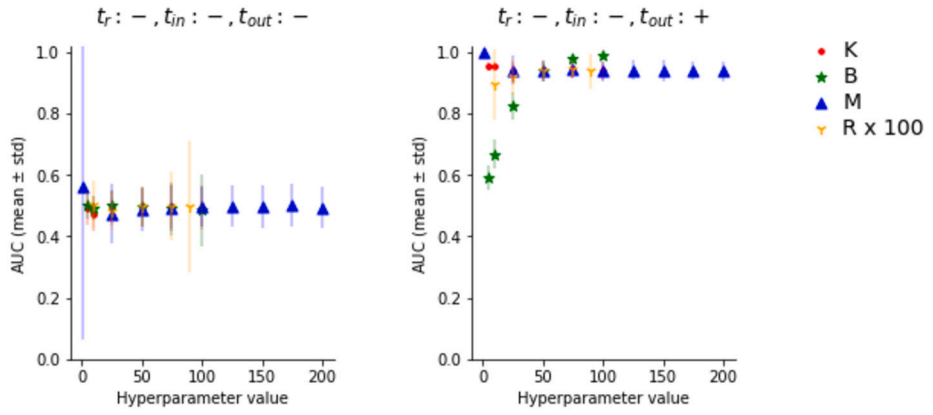


Fig. 7. Sensitivity analysis for the number of repetitions (K), the batch size (B), the number of batches (M) and the reference percentage (R), for $t_r = -$, $t_{in} = -$, $t_{out} = -$ (left) and $t_r = -$, $t_{in} = -$, $t_{out} = +$ (right).

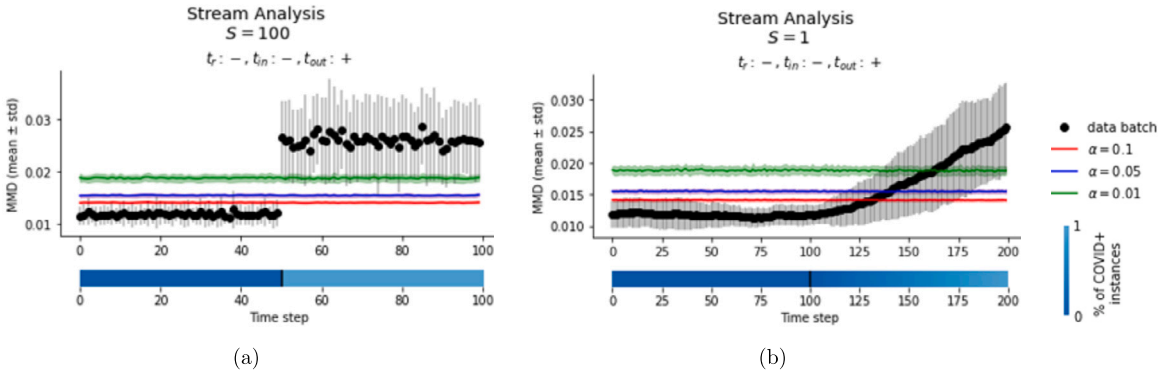


Fig. 8. Additional streaming simulations with strides 100 (a) and 1 (b). The lower bar shows the percentage of COVID+ patients in the data pool from which batches are being sampled (set_{in} during the first M steps and set_{out} thereafter).

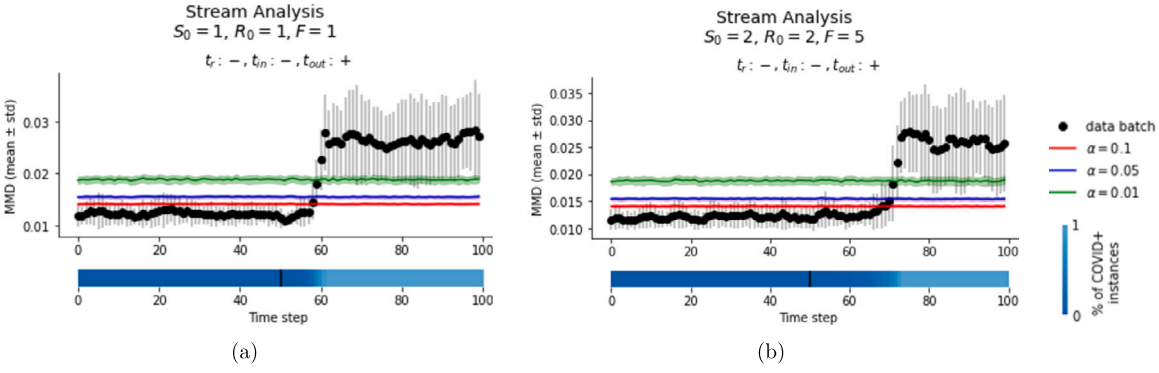


Fig. 9. Additional streaming simulations under exponential growth. The lower bar shows the percentage of COVID+ patients in the data pool from which batches are being sampled (set_{in} during the first M steps and set_{out} thereafter).

6. Discussion

In this section, we delve into the findings of our study and their implications. We discuss results, frame them into the context of existing literature and explore the limitations of our approach.

6.1. Instance-based approaches

The obtained results show that there are no relevant differences between the reconstruction loss of COVID− and COVID+ pneumonia scans (Table 4, AUC near 0.5), implying that this metric has little discriminative power in this problem. It could be hypothesized

that COVID− cases, being more similar to the autoencoder's training data, would exhibit a lower reconstruction loss. However, although some past studies were relatively successful using this approach [1,21], our study did not replicate these outcomes under our specific experimental conditions and design. Using Mahalanobis distance as an anomaly detector also failed to effectively separate COVID− and COVID+ cases. This contrasts with the results from [6], which motivated our use of this distance, where it is applied to varied features from multiple convolutional layers, targeting distinct anomalies like various body regions and pediatric scans. The differing contexts and anomaly types might explain why our application of this method produced different results. Together, these results suggest that instance-wise anomaly detection approaches are insufficient for distinguishing COVID-19 pneumonia cases from those of non-COVID-19 pneumonia.

6.2. Distributional approaches

Comparing COVID+ and COVID− test data to a COVID− reference set with MMD showed effective separation of the two types of data, yielding promising results (first line of Table 5). Additional results from the same table also show that when $t_r = -$, $t_{in} = -$, and $t_{out} = -$ the AUC is around 0.5, meaning that b_{in} and b_{out} are not distinguishable, which is expected, considering that they are made up of the same type of data. These results suggested that our distributional approach is effective in discriminating between sets of COVID− and COVID+ images, supporting our proposal of favoring it over instance-wise approaches for identifying COVID-19-related changes.

Early access to a substantial volume of labeled drifted data is often unfeasible since data are generated through time and labels will typically not be available. Moreover, in day-to-day practice, the unpredictable nature of emerging drifts, including those from new and unrelated diseases, underscores the value of keeping our system agnostic to specific drifts. The *stream* experiments were introduced to address these concerns, considering gradual data batch availability and no assumptions regarding the nature or labels of test data. Fig. 6 shows a clear separability in terms of MMD among batches generated from set_{in} (the first 50 time steps) and set_{out} (the last 50 time steps), especially after the transition phase (where the batches comprise a mixture of COVID− and COVID+ data), when data stops being sampled from set_{in} and starts to be sampled from set_{out}. Despite a relatively high standard deviation, there is minimal overlap between these two types of batches.

The sensitivity analyses show that results stabilize with increased repetitions K and the impact of batch size B on discriminative power, finding that smaller batches led to reduced discrimination, while larger batches increased standard deviation in some cases. This analysis provided valuable insights on the minimum viable number of COVID+ data samples essential for establishing a distinguishable distribution from that of COVID− scans. These insights guide us on how many samples we need to accumulate before conducting a new test to check for drift. Instead of using physical time, which can vary significantly between different clinical environments, we ought to reassess based on the number of CXRs accumulated. This approach offers a more universally applicable method because the rate at which CXRs are gathered differs widely between hospitals, making our strategy adaptable to varying operational workflows. The reference dataset's size has limited influence on the results, except in extreme scenarios where R is very small (it lacks representativeness) or R is too large (leaving too little data for testing). In the streaming analyses, a smaller stride results in smoother transitions when we start sampling from set_{out}, while the opposite happens for larger strides.

The existing COVID-19 classification and detection methods outlined in the literature predominantly adopt an instance-based rather than a distributional approach. As detailed in Section 2, some existing studies [8,28,30] fail to incorporate the negative cohort of the BIMCV-COVID19 dataset, which was released later than that of the positive cohort, and therefore select COVID+ and COVID− data from disparate sources, which may impact the reliability of the results. The study in [29] combines data from various sources, including BIMCV, for COVID-19 anomaly detection, achieving an AUC of less than 0.8. In [19], the same data as in [29] are used, and an AUC of 0.77 is reported. The results of our distributional approach are therefore competitive with the instance-wise methods reported in the literature. The most meaningful comparison between methodologies can be drawn against our baselines, given that we use the same data under the same conditions for all experiments.

6.2.1. Simulating an emerging pandemic

The discriminative ability of MMD drops as the drift purity (percentage of COVID+ cases in set_{out}) decreases. The distributions of set_{in} and set_{out} converge as purity decreases, making b_{in} and b_{out} batches more similar. The simulations in Fig. 6 depict a scenario where COVID+ instances emerge gradually and follow an exponential growth pattern, providing valuable insights into how early a drift of this nature could be detected. Notably, the detection of COVID-19-related changes appears to become feasible around 80% purity. Detecting subtle drifts is challenging, with discriminative power emerging only when more than half of set_{out} samples are COVID+ (Fig. 5). This is particularly relevant for COVID+ cases, where images aren't drastically different from $COVID-$ pneumonia cases, rendering point-wise comparisons ineffective. Detection timing will vary for different shift types, occurring earlier or later depending on the visual characteristics of the drift in the scans.

6.2.2. Calibration and transferability

The use of BIMCV-COVID19-PadChest as an external non-COVID pneumonia data source raised important calibration issues. When BIMCV-COVID19-PadChest was used for both t_r and t_{in} in Table 6, the MMD scores consistently indicate drift to both D^+ and D^- data due to the dataset shift between sources, which limited our ability to differentiate the BIMCV-COVID19 cohorts based on MMD values. However, when employing PadChest for set_r, COVID− data for set_{in}, and COVID+ data for set_{out}, the AUC is high (fourth line in Table 6), indicating that the MMD between the PadChest reference and data from D^- is smaller than the MMD between the PadChest reference and data from D^+. Using external data as a reference yields increased absolute values of the MMDs, but the AUC

evaluates class separability based on ranking, and not magnitudes. Therefore, a high AUC is maintained as data separability remains intact despite greater magnitudes.

The same patterns emerge in the stream simulations; using $t_r = p$ yields higher MMD scores (Fig. 6b) than using D^- as a reference (Fig. 6a), but proper calibration enables establishing an MMD threshold to distinguish between b_{in} and b_{out} batches. Collectively, these results suggest the need for system calibration during deployment. This is the main reason why in this work, we refrain from discussing a specific rule or threshold for defining a "drift alarm". Instead, we present a series of results stemming from multiple simulations and leave it up to the prospective user to define the most pertinent and effective approach for their specific context. The simulations presented serve as guiding benchmarks for the evolution of a prospective system grounded in this approach. Our findings demonstrate the applicability of our results to new data while emphasizing the importance of calibration with different data sources, such as data originating from a different healthcare facility.

6.3. Limitations

While our study offers valuable insights, it is focused exclusively on pneumonia cases. Although this problem is more complex than separating COVID-19 cases from normal CXRs, if this system were to be implemented in a fully automated manner, a previous model to identify CXRs showing signs of pneumonia should first be available. Our goal was to use pneumonia as a well-known reference disease and to identify deviations that might indicate the presence of a novel pathogen with analogous features. Another limitation worth considering is the potential impact of external factors, unrelated to the pathological features observed in the scans, on the data of the positive and negative cohorts. However, this should not be the case since both cohorts were collected similarly. If present, such distributional differences could also be viewed as information rather than as biases. If discernible patterns exist within these differences, capturing them may be desirable. One potential approach to address these uncertainties and inaccuracies in radiographic images could be the incorporation of fuzzy classification techniques. These techniques could merge similar images in a fuzzy sense, extracting a single representative image for each group [36]. Classification would then compare the similarity values between the images to be classified and the representative images of each group in a fuzzy sense.

7. Conclusions

In this study, we explored the challenge of distinguishing COVID-19 pneumonia cases from other pneumonia types using chest X-rays. We initially investigated instance-based out-of-distribution detection approaches, such as autoencoder-based reconstruction loss and Mahalanobis distance, but found that these methods were ineffective at discriminating between COVID-19 and non-COVID-19 cases, mostly due to the radiological similarities between different pneumonia types. We introduced a distributional approach, utilizing maximum mean discrepancy to compare populations of COVID-19 and non-COVID-19 cases. Our findings demonstrated the effectiveness of this populational approach in distinguishing COVID+ from COVID− cases, when compared to a COVID− reference set. Rather than targeting individual cases, our approach is designed to act as an early warning system, signaling noteworthy shifts within a population that may warrant further investigation. This approach showed promise in scenarios of gradual changes in the data distribution occur, revealing its potential in monitoring and flagging emerging pandemics with evolving characteristics. However, we emphasize the importance of calibration when using external reference datasets, as dataset shifts can affect the absolute maximum mean discrepancy values. Calibration becomes especially crucial when deploying the method across different healthcare settings.

Although our study has some limitations, it provides valuable insights into the challenges and potential solutions for improving the accuracy of identifying COVID-19-related changes in chest radiographs. Ultimately, this research offers a framework for addressing similar problems and monitoring changes in radiographic manifestations across various clinical scenarios in the future, such as the anticipation of a new pandemic caused by a pathogen whose infection causes lung-related manifestations.

Building on the success of the proposed distributional approach, future research should explore several avenues to address its limitations and enhance its practical application. More specifically, the implementation of these methods in real-world healthcare settings should be assessed. Adapting and testing the methodology across diverse clinical settings and various imaging modalities, such as CT scans, would be beneficial. The integration of multimodal data, including clinical parameters like patient history and demographic details, could potentially improve the results, particularly in detecting early-stage changes. Further exploration of alternative metrics beyond the maximum mean discrepancy could also contribute to enhancing the framework. Moreover, we could also explore the use of incremental model updates to manage changes in data, which could improve the sensitivity of the detection system. Collaborative efforts with clinical practitioners will be crucial to tailor the model's development to real-world applicability and to ensure that it effectively complements existing diagnostic tools.

Ethics approval and consent to participate

Publicly available data was used.

CRediT authorship contribution statement

Sofia C. Pereira: Writing – original draft, Methodology, Investigation, Conceptualization. **Joana Rocha:** Writing – review & editing, Conceptualization. **Aurélio Campilho:** Writing – review & editing, Supervision. **Ana Maria Mendonça:** Writing – review & editing, Supervision.

Declaration of competing interest

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

Data availability

Publicly available data was used. It is accessible via https://bimcv.cipf.es/bimcv-projects/bimcv-covid19/.

Acknowledgements

This work was supported by the European Regional Development Fund and by National Funds through the FCT - Portuguese Foundation for Science and Technology, I.P. within the scope of the CMU Portugal Program (NORTE-01-0247-FEDER-045905) and LA/P/0063/2020. The work of S. C. P. was supported by the FCT grant contract 2020.10169.BD. The work of J. R. was supported by the FCT grant contract 2020.06595.BD.

Appendix A. Preprocessing of the BIMCV-COVID19 dataset

As mentioned in Section 3, we started by manually discarding images that were missing metadata, corrupted, from wrong body parts, of insufficient quality, or from pediatric patients. After creating the dataset, its authors found a set of patients from the negative cohort that were also in the positive cohort. They realized that these were false positive instances wrongly examined in the early stages of the pandemic, and later released a list of the common subjects (i.e., a subject with two IDs, one in each cohort). All scans from patients in this list were also discarded. For the images included in the study, we first performed histogram-based contrast enhancement, to improve under/over-exposed images. Then, we selected the region of interest by detecting and removing (when present) black/white framings around relevant image data. Finally, the images were normalized to the 0-255 range. After manual inspection, we found that the information on the images' monochrome is sometimes incorrect, resulting in inverted images. To address this, an automatic inversion classification CNN identified and corrected inverted images. The model was trained on the BIMCV-COVID19-PADCHEST dataset, where half the images were randomly inverted, using a 90/10 train/test split. It trained for 10 epochs using ADAM [22] and a learning rate of 10^{-4} on an NVIDIA GTX 1080 GPU.

References

[1] F.F. Abir, M.E. Chowdhury, M.I. Tapotee, A. Mushtak, A. Khandakar, S. Mahmud, A. Hasan, PCovNet+: a CNN-VAE anomaly detection framework with LSTM embeddings for smartwatch-based COVID-19 detection, Eng. Appl. Artif. Intell. 122 (2023) 106130, https://doi.org/10.1016/j.engappai.2023.106130.

[2] A. Albiol, F. Albiol, R. Paredes, J.M. Plasencia-Martínez, A.B. Barrio, J.M. Santos, S. Tortajada, V.M.G. Montaño, C.E.R. Godoy, S.F. Gómez, E. Oliver-Garcia, M. de la Iglesia Vayá, F.L.M. Pérez, J.I.R. Madrid, A comparison of Covid-19 early detection between convolutional neural networks and radiologists, Insights Imaging 13 (2022) 1–12, https://doi.org/10.1186/S13244-022-01250-3.

[3] S.U. Amin, S. Taj, A. Hussain, S. Seo, An automated chest X-ray analysis for covid-19, tuberculosis, and pneumonia employing ensemble learning approach, Biomed. Signal Process. Control 87 (2024) 105408, https://doi.org/10.1016/j.bspc.2023.105408.

[4] K.R. Bhatele, A. Jha, D. Tiwari, M. Bhatele, S. Sharma, M.R. Mithora, S. Singhal, Covid-19 detection: a systematic review of machine and deep learning-based approaches utilizing chest X-rays and ct scans, Cogn. Comput. 1 (2022) 1–38, https://doi.org/10.1007/S12559-022-10076-6.

[5] A. Bustos, A. Pertusa, J.M. Salinas, M. de la Iglesia-Vayá, PadChest: a large chest X-ray image dataset with multi-label annotated reports, Med. Image Anal. 66 (2020) 101797, https://doi.org/10.1016/j.media.2020.101797.

[6] E. Calli, B. Van Ginneken, E. Sogancioglu, K. Murphy, Frodo: an in-depth analysis of a system to reject outlier samples from a trained neural network, IEEE Trans. Med. Imaging 42 (2023) 971–981, https://doi.org/10.1109/TMI.2022.3221898.

[7] W. Chen, Z. Li, J. Guo, Domain adaptation learning based on structural similarity weighted mean discrepancy for credit risk classification, IEEE Intell. Syst. 35 (2020) 41–51, https://doi.org/10.1109/MIS.2020.2972791.

[8] M. Chetoui, M.A. Akhloufi, Deep efficient neural networks for explainable COVID-19 detection on CXR images, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 12798 LNAI, 2021, pp. 329–340.

[9] J. Cleverley, J. Piper, M.M. Jones, The role of chest radiography in confirming covid-19 pneumonia, BMJ 370 (2020), https://doi.org/10.1136/bmj.m2426.

[10] J.P. Cohen, J.D. Viviano, P. Bertin, P. Morrison, P. Torabian, M. Guarrera, M.P. Lungren, A. Chaudhari, R. Brooks, M. Hashir, H. Bertrand, Torchxrayvision: a library of chest X-ray datasets and models, in: E. Konukoglu, B. Menze, A. Venkataraman, C. Baumgartner, Q. Dou, S. Albarqouni (Eds.), Proceedings of the 5th International Conference on Medical Imaging with Deep Learning, 2022, pp. 231–249, https://proceedings.mlr.press/v172/cohen22a.html. (Accessed 5 January 2023).

[11] B.G.S. Cruz, M.N. Bossa, J. Sölter, A.D. Husch, Public Covid-19 X-ray datasets and their impact on model bias – a systematic review of a significant problem, Med. Image Anal. 74 (2021) 102225, https://doi.org/10.1016/J.MEDIA.2021.102225.

[12] M.J. Gangeh, H. Tadayyon, L. Sannachi, A. Sadeghi-Naini, W.T. Tran, G.J. Czarnota, Computer aided theragnosis using quantitative ultrasound spectroscopy and maximum mean discrepancy in locally advanced breast cancer, IEEE Trans. Med. Imaging 35 (2016) 778–790, https://doi.org/10.1109/TMI.2015.2495246.

[13] A. Gretton, K.M. Borgwardt, M.J. Rasch, B. Schölkopf, A. Smola, A kernel two-sample test, J. Mach. Learn. Res. 13 (2012) 723–773, http://jmlr.org/papers/v13/gretton12a.html. (Accessed 5 January 2023).

[14] X. Guo, J.W. Gichoya, H. Trivedi, S. Purkayastha, I. Banerjee, Medshift: automated identification of shift data for medical image dataset curation, IEEE J. Biomed. Health Inform. 27 (2023) 3936–3947, https://doi.org/10.1109/JBHI.2023.3275104.

[15] J. Hu, X. Gu, Z. Wang, X. Gu, Active consistency network for multi-source domain generalization in brain tumor segmentation, Biomed. Signal Process. Control 86 (2023) 105132, https://doi.org/10.1016/j.bspc.2023.105132.

[16] M. de la Iglesia Vayá, J.M. Saborit-Torres, J.A. Montell Serrano, E. Oliver-Garcia, A. Pertusa, A. Bustos, M. Cazorla, J. Galant, X. Barber, D. Orozco-Beltrán, F. García-García, M. Caparrós, G. González, J.M. Salinas, BIMCV COVID-19-: a large annotated dataset of RX and CT images from COVID-19 patients, IEEE Dataport (2021), https://doi.org/10.21227/m4j2-ap59.

[17] M. de la Iglesia Vayá, J.M. Saborit-Torres, J.A. Montell Serrano, E. Oliver-Garcia, A. Pertusa, A. Bustos, M. Cazorla, J. Galant, X. Barber, D. Orozco-Beltrán, F. García-García, M. Caparrós, G. González, J.M. Salinas, BIMCV COVID-19+: a large annotated dataset of RX and CT images from COVID-19 patients, IEEE Dataport (2021), https://doi.org/10.21227/w3aw-rv39.

[18] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, J. Seekins, D.A. Mong, S.S. Halabi, J.K. Sandberg, R. Jones, D.B. Larson, C.P. Langlotz, B.N. Patel, M.P. Lungren, A.Y. Ng, CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison, in: 33rd AAAI Conference on Artificial Intelligence, 2019, pp. 590–597.

[19] N. Jahan, M.A.M. Hasan, Autoencoder-based unsupervised anomaly detection for Covid-19 screening on chest X-ray images, in: 2022 19th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE), 2022, pp. 1–6.

[20] A.E. Johnson, T.J. Pollard, S.J. Berkowitz, N.R. Greenbaum, M.P. Lungren, Chih-ying Deng, R.G. Mark, S. Horng, MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports, Sci. Data 6 (2019) 1–8, https://doi.org/10.1038/s41597-019-0322-0.

[21] Y. Karadayi, M.N. Aydin, A.S. Öğrencí, Unsupervised anomaly detection in multivariate spatio-temporal data using deep learning: early detection of COVID-19 outbreak in Italy, IEEE Access 8 (2020) 164155–164177, https://doi.org/10.1109/ACCESS.2020.3022366.

[22] D.P. Kingma, J.L. Ba, Adam: a method for stochastic optimization, in: International Conference on Learning Representations, ICLR, 2015, https://arxiv.org/abs/1412.6980v9. (Accessed 5 January 2023).

[23] L.M. Koch, C.M. Schürch, A. Gretton, P. Berens, Hidden in plain sight: subgroup shifts escape OOD detection, in: E. Konukoglu, B. Menze, A. Venkataraman, C. Baumgartner, Q. Dou, S. Albarqouni (Eds.), Proceedings of the 5th International Conference on Medical Imaging with Deep Learning, PMLR, 2022, pp. 726–740, https://proceedings.mlr.press/v172/koch22a.html. (Accessed 5 January 2023).

[24] P.C. Mahalanobis, On the generalized distance in statistics, Proc. Natl. Inst. Sci. (Calcutta) 2 (1936) 49–55, https://link.springer.com/article/10.1007/s13171-019-00164-5. (Accessed 5 January 2023).

[25] E. Martínez Chamorro, A. Díez Tascón, L. Ibáñez Sanz, S. Ossaba Vélez, S. Borruel Nacenta, Radiologic diagnosis of patients with covid-19, Radiología (English Edition) 63 (2021) 56–73, https://doi.org/10.1016/j.rxeng.2020.11.001, https://www.sciencedirect.com/science/article/pii/S2173510721000033.

[26] D. Meedeniya, H. Kumarasinghe, S. Kolonne, C. Fernando, I.D. la Torre Díez, G. Marques, Chest X-ray analysis empowered with deep learning: a systematic review, Appl. Soft Comput. 126 (2022) 109319, https://doi.org/10.1016/J.ASOC.2022.109319.

[27] T. Mishra, M. Wang, A.A. Metwally, G.K. Bogu, A.W. Brooks, A. Bahmani, A. Alavi, A. Celli, E. Higgs, O. Dagan-Rosenfeld, B. Fay, S. Kirkpatrick, R. Kellogg, M. Gibson, T. Wang, E.M. Hunting, P. Mamic, A.B. Ganz, B. Rolnik, X. Li, M.P. Snyder, Pre-symptomatic detection of COVID-19 from smartwatch data, Nat. Biomed. Eng. 2020 (2020) 1208–1220, https://doi.org/10.1038/s41551-020-00640-6.

[28] A. Miyazaki, K. Ikejima, M. Nishio, M. Yabuta, H. Matsuo, K. Onoue, T. Matsunaga, E. Nishioka, A. Kono, D. Yamada, K. Oba, R. Ishikura, T. Murakami, Computer-aided diagnosis of chest X-ray for COVID-19 diagnosis in external validation study by radiologists with and without deep learning system, Sci. Rep. 13 (2023) 17533, https://doi.org/10.1038/s41598-023-44818-9.

[29] S. Motamed, P. Rogalla, F. Khalvati, RANDGAN: randomized generative adversarial network for detection of COVID-19 in chest X-ray, Sci. Rep. 11 (2021) 8602, https://doi.org/10.1038/s41598-021-87994-2.

[30] J. Pedrosa, G. Aresta, C. Ferreira, C. Carvalho, J. Silva, P. Sousa, L. Ribeiro, A.M. Mendonça, A. Campilho, Assessing clinical applicability of COVID-19 detection in chest radiography with deep learning, Sci. Rep. 2022 (2022) 1–17, https://doi.org/10.1038/s41598-022-10568-3.

[31] S. Rabanser, S. Günnemann, Z.C. Lipton, Failing Loudly: An Empirical Study of Methods for Detecting Dataset Shift, Curran Associates Inc., Red Hook, NY, USA, 2019, https://dl.acm.org/doi/abs/10.5555/3454287.3454412. (Accessed 5 January 2023).

[32] M. Roberts, D. Driggs, M. Thorpe, J. Gilbey, M. Yeung, S. Ursprung, A.I. Aviles-Rivero, C. Etmann, C. McCague, L. Beer, J.R. Weir-McCall, Z. Teng, E. Gkrania-Klotsas, A. Ruggiero, A. Korhonen, E. Jefferson, E. Ako, G. Langs, G. Gozaliasl, G. Yang, H. Prosch, J. Preller, J. Stanczuk, J. Tang, J. Hofmanninger, J. Babar, L.E. Sánchez, M. Thillai, P.M. Gonzalez, P. Teare, X. Zhu, M. Patel, C. Cafolla, H. Azadbakht, J. Jacob, J. Lowe, K. Zhang, K. Bradley, M. Wassin, M. Holzer, K. Ji, M.D. Ortet, T. Ai, N. Walton, P. Lio, S. Stranks, T. Shadbahr, W. Lin, Y. Zha, Z. Niu, J.H.F. Rudd, E. Sala, C.B. Schönlieb, Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans, Nat. Mach. Intell. (2021), https://doi.org/10.1038/s42256-021-00307-0.

[33] A. Soin, J. Merkow, J. Long, J.P. Cohen, S. Saligrama, S. Kaiser, S. Borg, I. Tarapov, M.P. Lungren, CheXstray: real-time multi-modal data concordance for drift detection in medical imaging AI, https://doi.org/10.48550/arxiv.2202.02833, 2022.

[34] N. Subramanian, O. Elharrouss, S. Al-Maadeed, M. Chowdhury, A review of deep learning-based detection methods for covid-19, Comput. Biol. Med. 143 (2022) 105233, https://doi.org/10.1016/j.compbiomed.2022.105233.

[35] M. Surya Bhupal Rao, Y. Mallikarjuna Rao, C. Venkataiah, G. Murthy, M. Dharani, M. Jayamma, Deep learning based classification of covid-19 severity using hierarchical deep maxout model, Biomed. Signal Process. Control 88 (2024) 105653, https://doi.org/10.1016/j.bspc.2023.105653.

[36] M. Versaci, G. Angiulli, P. Crucitti, D. De Carlo, F. Laganà, D. Pellicanò, A. Palumbo, A fuzzy similarity-based approach to classify numerically simulated and experimentally detected carbon fiber-reinforced polymer plate defects, Sensors 22 (2022), https://doi.org/10.3390/s22114232.

[37] T. Viehmann, L. Antiga, D. Cortinovis, L. Lozza, TorchDrift: drift detection for PyTorch, https://github.com/TorchDrift/TorchDrift. (Accessed 5 January 2023), 2020.

[38] W. Wang, Y. Xu, R. Gao, R. Lu, K. Han, G. Wu, W. Tan, Detection of SARS-CoV-2 in different types of clinical specimens, JAMA J. Am. Med. Assoc. 323 (2020) 1843–1844, https://doi.org/10.1001/jama.2020.3786.

[39] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, R.M. Summers, ChestX-ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, in: Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, 2017, pp. 3462–3471.

[40] J. Wei, G. Wang, Fine-grained out-of-distribution detection of medical images using combination of feature uncertainty and Mahalanobis distance, in: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI), 2023, pp. 1–5.

[41] J. Yanase, E. Triantaphyllou, A systematic survey of computer-aided diagnosis in medicine: past and present developments, Expert Syst. Appl. 138 (2019), https://doi.org/10.1016/j.eswa.2019.112821.

[42] F. lah Yassaanah Issahaku, X. Liu, K. Lu, X. Fang, S.B. Danwana, E. Asimeng, Multimodal deep learning model for covid-19 detection, Biomed. Signal Process. Control 91 (2024) 105906, https://doi.org/10.1016/j.bspc.2023.105906.

[43] Z. Zhao, H. Peng, X. Zhang, Y. Zheng, F. Chen, L. Fang, J. Li, Identification of lung cancer gene markers through kernel maximum mean discrepancy and information entropy, BMC Med. Genom. 12 (2019) 1–10, https://doi.org/10.1186/S12920-019-0630-4/TABLES/5.