PrepNet: A Convolutional Auto-Encoder to Homogenize CT Scans for Cross-Dataset Medical Image Analysis

Mohammadreza Amirian*†, Javier A. Montoya-Zegarra*, Jonathan Gruss*¶, Yves D. Stebler*¶,
Ahmet Selman Bozkir*¶, Marco Calandri‡, Friedhelm Schwenker† and Thilo Stadelmann*§
*ZHAW School of Engineering, 8400 Winterthur, Switzerland
†Ulm University, Institute of Neural Information Processing, 89081 Ulm, Germany
‡University of Turin, Department of Oncology, 10124 Turin, Italy
§Fellow, ECLT European Centre for Living Technology, 30123 Venice, Italy
¶These authors contributed equally to this work
amir@zhaw.ch

***Abstract*—With the spread of COVID-19 over the world, the need arose for fast and precise automatic triage mechanisms to decelerate the spread of the disease by reducing human efforts e.g. for image-based diagnosis. Although the literature has shown promising efforts in this direction, reported results do not consider the variability of CT scans acquired under varying circumstances, thus rendering resulting models unfit for use on data acquired using e.g. different scanner technologies. While COVID-19 diagnosis can now be done efficiently using PCR tests, this use case exemplifies the need for a methodology to overcome data variability issues in order to make medical image analysis models more widely applicable. In this paper, we explicitly address the variability issue using the example of COVID-19 diagnosis and propose a novel generative approach that aims at erasing the differences induced by e.g. the imaging technology while simultaneously introducing minimal changes to the CT scans through leveraging the idea of deep auto-encoders. The proposed prepossessing architecture (*PrepNet*) (i) is jointly trained on multiple CT scan datasets and (ii) is capable of extracting improved discriminative features for improved diagnosis. Experimental results on three public datasets (SARS-COVID-2, UCSD COVID-CT, MosMed) show that our model improves cross-dataset generalization by up to 11.84 percentage points despite a minor drop in within dataset performance.**

***Index Terms*—Adaptive preprocessing, domain adaptation, auto-encoder**

I. Introduction

A major challenge in rolling out machine learned models to a broad user base is the variability of data encountered in the real world. Models can only be expected to work well on data of similar distribution as has been used for training, but ubiquitously, differences in image acquisition setup hinder the applicability of a once developed model in novel settings. A recent example for the negative effects of such failure to adapt between different domains has been given at the start of the COVID-19 pandemic:

As of 2nd February 2021, this disease has caused over 100 million infections worldwide and over 2 million deaths according to the World Health Organisation (WHO) [1]. To alleviate this, rapid diagnosis of COVID-19 cases has been proven to be effective for decelerating the spread of the disease [2]. According to [2], [3], reverse transcriptase quantitative polymerase chain reaction (RT-qPCR) tests are accepted as the gold standard for the identification of positive cases. However, this type of test was not available in sufficient numbers at the beginning of the pandemic. Further, beyond being time-consuming, it relies on both human effort and expert knowledge. Thus, there arose a need for automatic diagnostic methods that can assist experts and reduce human efforts by targeting the automatic identification of COVID-19 positive cases. The literature has shown promising efforts in the automatic identification of COVID-19 cases from lung computed tomography (CT) scans using computer vision methods [4], [5], [6], [7]. Lessmann et al. addressed cross-vendor analysis (between different CT scanners such as Varian, Siemens, GE Healthcare, Philips and Canon) for 3D CT scans successfully [8]. However, it is demonstrated that a considerable drop in cross-dataset performance appears for the diagnosis of 2D CT scans acquired via different devices. Thus, the previously mentioned *within dataset variability* has the potential to discourage the community to merge and annotate data from multiple sources. As a result, combining datasets is a challenge posed not only for COVID detection but also for other applications in diagnosis and segmentation.

In this paper, we address domain adaptation of medical image analysis methods by proposing a deep convolutional neural network (CNN) for preprocessing 2D CT scans such that it is trained to fool a classifier that discriminates between various CT datasets, thus aiming to remove the within dataset variability. We evaluate the performance of the suggested method on the exemplary use case of predicting COVID-19 positive cases, due to the global variability in respective datasets and the availability of plenty of opportunities to compare. It should be noted that, the methodology is inspired by generative adversarial learning [9], [10]. Our contribution is twofold: (i) we propose a novel trainable preprocessing CNN architecture with a dual training objective that is capable of equalizing the variability of different CT-scanner technologies in the image domain as a pre-processor (*PrepNet*); (*ii*) we validate this model by showing the transferability of its diagnostic capabilities between different CT data sources based on common public benchmarks. We conduct experiments on the *SARS-CoV-2 CT-scan dataset* [11] and the *UCSD COVID-CT dataset* [12] as well as *MosMed dataset* [13]. Our results show that our *PrepNet* model improves the cross-dataset COVID-19 diagnosis performance (i.e., training on one dataset and testing on another) by 11.84 percentage points (pp) through creating a unified representation of multi-dataset CT scans.

II. Related Work

With the emergence of COVID-19, many studies and datasets have been proposed in the literature which show an increase in data diversity over time and the extent of related computer vision methods to deal with it [14], [15]. Horry et al. [2] utilize a transfer learning scheme to build various COVID-19 classifiers based on several off the shelf CNN models such as VGG16/19 [16], Resnet50 [17], InceptionV3 [18], Xception [19], and InceptionResnet [20]. They compared the generalization capability of various images sources such as X-ray, CT and ultrasound images and developed a pre-processing scheme for X-ray images to reduce noise at non-lung areas in order to decrease the effect of quality imbalance among the employed images. A VGG19 [16] coupled with ultrasound images is found to yield the best validation accuracy of 99%, while 84% have been achieved using CT scans [21].

He et al. [21] propose a sample-efficient learning concept called "Self-Trans" via synergetically combining transfer learning and contrastive self-supervised learning. They seek intrinsic visual patterns in CT scans without relying on labels created with human effort. Besides, they open-sourced their CT dataset involving 349 COVID-19 positive patients and 397 COVID-19 negatives [12]. They achieve an accuracy of 86% through unbiased feature representations together with a reduction of overfitting.

Mobiny et al. [22] propose the DECAPS approach with following contributions: (*i*) inverted dynamic routing [23] to avoid seeking visual features from non-related regions, (*ii*) training with a two-stage patch crop and drop strategy to encourage the network to focus on the useful areas, (*iii*) employing conditional generative adversarial networks for data augmentation. Experiments result 84.3% precision and 91.5% recall along with 87.6% accuracy. They additionally report results for the conventional deep classifiers DenseNet121 [24] and Resnet50 [17], yielding 82.5% and 80.8% accuracy, respectively. In contrast to this study, Pham [25] points out the negative impact of data augmentation in the context of CT-based COVID-19 image classification. In his study, the author fine-tunes various well-known pre-trained CNN models ranging from AlexNet [26] to NasNet-Large [27]. Experiments conducted on the already introduced CT dataset [12] credit a DenseNet-201 with the best accuracy of 96.2%. However, data augmentation using random vertical/horizontal flips (p=0.5), vertical/horizontal translation (±30 pixels) and scaling (±10%) yields a 6% accuracy drop on average.

Chaganti et al. [28] suggest a deep-reinforcement-learning-based scheme focusing on seeking doubtful lung areas on CT scans to localize abnormal portions. A recent study by [15], a novel architecture called "COVID-Net- CT-2" which utilizes machine-driven design exploration based on iterative constrained optimization is proposed [29]. The authors point out that one of the subtle problems of earlier studies is the limited number of patients and poor diversity of CT scans in terms of multi-nationality. Therefore, they introduce the two large-scale COVID-19 CT datasets called "COVIDx CT-2A" and "COVIDx CT-2B" gathered from $4,501$ patients from at least 15 countries, totally comprising 194.922 and 201.103 images respectively. Experiments show that the architecture achieves a sensitivity of 99.0% and an accuracy of 98.1%, which competes with radiologist-level decision making capability. The study deals with variability in the patients' ethnicity, while CT scans generated by various vendors' devices exhibit visual differences, artifacts, and variable intensities that are never addressed so far. Thus, independent from the reported success of some deep learning architecture, it is likely to witness a drop in prediction accuracy during inference when a test image is taken with a different device as has been used for training. Motivated by this issue, we propose to employ a pre-processing network (*PrepNet*) to standardize CT images with respect to the visual differences among datasets prior to training of any final diagnosis model, relying on generative architectures since they showed very promising results for similar tasks [22]. An advantage of this approach is that the *PrepNet* can be combined with any downstream diagnosis model, thus leveraging future progress there without additional costs while improving cross-dataset performance.

Two research papers closely related to the goal of domain adaptation in this study are presented by Lessmann et al. addressing cross-vendor diagnosis [8] and Amyar et al. using auto-encoders in multi-task learning [30]. Nevertheless, Lessmann et al. did not confront a considerable cross-vendor performance drop because of using a richer source of information (3D scans) as explained in [31]. Amyar et al. leveraged multi-task learning and trained an auto-encoder besides a segmentation and classification model for COVID-19 diagnosis. However, they did not aim at removing the cross-dataset variability of the scans. This study focuses on homogenizing the 2D CT scans by reducing cross-dataset information.

III. Methodology

In this section, we give details of our *PrepNet* model in terms of network architecture, core modules, and loss functions. The architecture of our proposed model is presented in Figure 1. For a group of $\mathcal{N}$ input CT scans $\{\mathcal{X}^n\}_{n=1}^{\mathcal{N}}$, coming from different CT vendors' devices, our model extracts multi-scale discriminative feature maps through an auto-encoder and reconstructs the original CT scans $\{\hat{\mathcal{X}}^n\}_{n=1}^{\mathcal{N}}$. The reconstructed CT scans are next fed into a dataset/technology

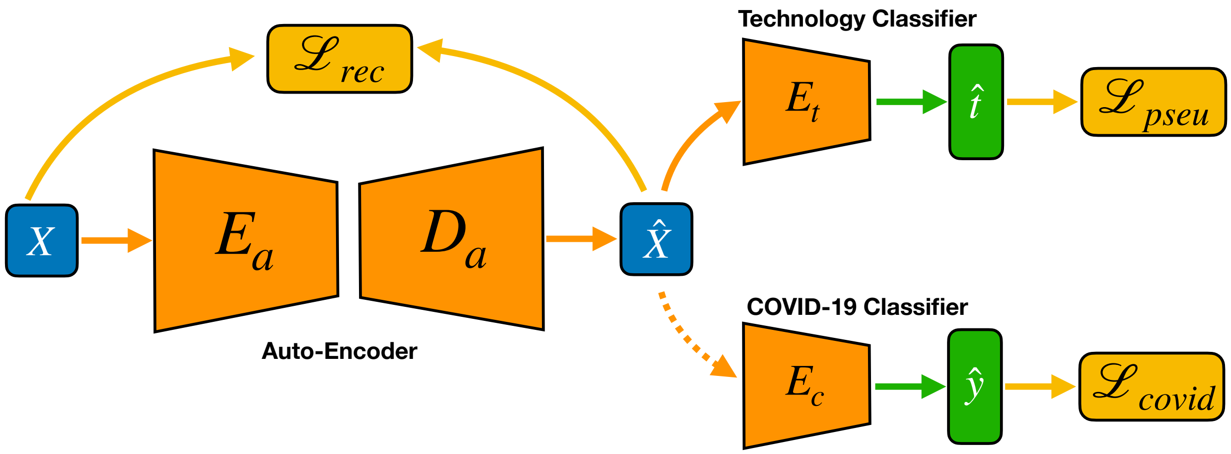

Fig. 1. The architecture of our proposed *PrepNet* model consists of three main modules: (i) an auto-encoder that acts as a CT cross-dataset homogenizer; (ii) a multi CT-technology classifier; and (iii) a COVID-19 binary classifier. The auto-encoder and the multi CT technology classifier are trained adversarially. The binary COVID-19 classifier is independently trained using the auto-encoder's output.

classification branch which acts as a pseudo-label classifier and is responsible for discriminating among different CT datasets. Once this model is trained end-to-end in an adversarial way, the reconstructed CT scans are fed into a COVID-19 classifier which is trained directly on the reconstructed CT-scans. The COVID-19 classification branch is responsible for the classification of healthy vs. non-healthy patients. The complete network model with its main modules are described in more detail below.

A. Model Architecture

Auto-Encoder Module: We feed a CT scan image $\mathcal{X}^n$ into our auto-encoder (E_a and D_a) and obtain a reconstructed version $\hat{\mathcal{X}}^n$ given by $\hat{\mathcal{X}}^n = D_a(E_a(\mathcal{X}^n))$. The encoder E_a is based on the standard classification network VGG-Net [16], whilst the decoder D_a is a convolutional network with the same number of layers as the encoder. We add skip-connections from E_a to D_a to recover the spatial information lost during the down-sampling operations.

Dataset Classifier Module: The CT dataset classifier E_t receives the reconstructed CT scan $\hat{\mathcal{X}}^n$ from the auto-encoder as input and feeds it into an encoder branch $E_t(\hat{\mathcal{X}}^n)$ that classifies the CT dataset/technology. In our experiments, E_t relies on the VGG-Net architecture as well.

COVID-19 Classifier Module: The COVID-19 classifier E_c is also uses several backbone architectures. Given a reconstructed CT scan $\hat{\mathcal{X}}^n$, it outputs COVID vs. non-COVID predictions, i.e. $E_c(\hat{\mathcal{X}}^n)$.

B. Loss Functions and Evaluation Metric

The complete loss function of *PrepNet* is based on the various terms presented in Figure 1. It comprises a reconstruction loss $\mathcal{L}_{rec}$ and two classification losses $\mathcal{L}_{pseu}$ and $\mathcal{L}_{covid}$:

$$\mathcal{L}_{total} = \mathcal{L}_{rec} + \mathcal{L}_{pseu} + \mathcal{L}_{covid} \tag{1}$$

Given the labeled dataset $\mathcal{D} = \{(\mathcal{X}^n, y^n, p^n)\}_n^N$ comprising the CT scans $\mathcal{X}^n$ together with their binary COVID label y^n and the CT-dataset pseudo label p^n, the auto-encoder reconstruction loss is given by $\mathcal{L}_{rec} = \sum_n \|\mathcal{X}^n - \hat{\mathcal{X}}^n\|_2^2$; the COVID-19 binary classification loss is denoted $\mathcal{L}_{covid} = -\sum_n y_n \log \hat{y}_n + (1-y_n)\log(1-\hat{y}_n)$; the CT dataset pseudo label is computed by $\mathcal{L}_{pseu} = -\sum_n p_n \log \hat{p}_n$.

To measure the COVID-19 detection performance and to minimize the effect of class imbalance in datasets, we use the balanced accuracy metric (BA) [32]

$$BA = \frac{TP}{P} + \frac{TN}{N} \tag{2}$$

where P and N are the number of positive and negative samples respectively and TP and TN denote the number of true positive and true negative predictions, respectively. In addition, we also use specificity, sensitivity, and area under the curve to evaluate the COVID-19 performance results.

IV. Experiments

A. Datasets

We use three public datasets to validate our approach experimentally. The *SARS-CoV-2 CT-scan dataset* [11] comprises a total of $4,173$ CT images of real patients from the Public Hospital of the Government Employees of Sao Paulo (HSPM) and the Metropolitan Hospital of Lapa, both in Sao Paulo - Brazil ($2,168$ positive/infected and 768 healthy patients). Moreover, $1,247$ CT scans belong to patients who have other pulmonary diseases. The CT image annotations (positive vs. negative) have been done by three different clinicians. Note that during our visual inspection we found two erroneous images (i.e. unrelated to the problem domain) and excluded them from the dataset. In addition, we also excluded the $1,247$ pulmonary diseased patients.

The *UCSD COVID-CT dataset* [12] has been collected in the Tongji Hospital in Wuhan, China during the outbreak of COVID-19 between the months of January/2020 and April/2020. This dataset contains 349 CT images from infected patients and 397 from non-infected patients. All images have been annotated by a senior radiologist of the same hospital. As reported by [22], heights of the images in this dataset range between 153 and $1,853$ pixels with an average of 491 pixels, whereas the widths vary between 124 and $1,458$ pixels (average of 383 pixels). For partitioning, we follow the splitting guideline provided by the authors of the dataset. Table I summarizes the train, validation and test splits for each dataset.

The *MosMed dataset* [13] was collected by the Moscow Health Care Department from different municipal hospitals in Russia between March/2020 and April/2020. The dataset contains axial CT images from 1110 patients with different levels of COVID-19 severity, ranging from mild to critical cases and also healthy patients. Some image samples of each dataset are provided in Figure 2.

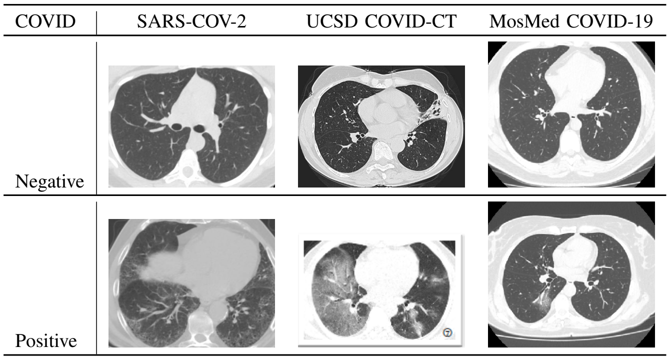

Fig. 2. COVID-19 positive and negative samples for each used dataset. Note the variabilities in terms of texture, size, and shape across datasets.

B. Implementation Details

We run all our experiments using the publicly available Pytorch 1.5.0 library and an NVIDIA VP100 GPU (32 GB of VRAM). During network training, each image is first resized according to the input size of the classifiers' backbones; we use histogram equalization as a fixed preprocessing step, then apply the mean and standard deviation of ImageNet pretrained models. We train *PrepNet* using the AdamW optimizer [33]. We perform a 24 hour hyperparameter search with six parallel runs using the Bayesian search strategy with Hyperband for early-stopping on one GPUs [34]. The hyperparameter search improves the chance of avoiding local minima and presenting optimal results of every configuration. The best model is selected based on the optimal validation performances. During training, we first train the auto-encoder for 20 epochs and warm up the dataset classification branch for 2 epochs before we start with the adversarial training. Once the adversarial training is finished, we train the COVID classification branch independently from the other two branches using the output of the auto-encoder/*PrepNet*.

C. Experimental Results

The within- and cross-dataset performance of the proposed preprocessing schemes are presented in Table II. In order to observe possible overfitting, we report the hold out test set performance on each dataset. The cross-dataset performance is evaluated by measuring the balanced accuracy (minimizing the effect of class imbalance) of the models trained on one dataset and tested on the other. We report results using the balanced accuracy of the models trained on the SARS-COV-2 and UCSD COVID-CT datasets. Further metrics also include sensitivity (Sens), specificity (Spec) and area under the curve (AUC). In the rows, we present the datasets used during training. Furthermore, we group the results by model. The first group of results are related to the COVID classifier (VGG-19 pre-trained model), that is trained and evaluated on the original CT scans. The second group of results is related to the auto-encoder alone trained on both datasets in a self-supervised manner to minimize the reconstruction loss. The third group of results relate to full *PrepNet* preprocessing before training the classifiers.

The results in Table II show that the average cross-dataset performance (over all dataset splits) of models trained on original data increases by 6.77pp after using the pure auto-encoder model, and by 11.84pp through *PrepNet*. However, the average test accuracy for within-dataset evaluation declines by 0.32pp and 1.83pp after applying the baseline auto-encoder or *PrepNet*, respectively. A discussion regarding this effect is presented in the next section.

In our experiments, we use the VGG19 [16] as the baseline model because it is more straight-forward to train and has shown good generalization properties on 2D medical images based on previous practical experiments[1]. Besides that, the VGG architecture has been also successfully applied for COVID-19 identification [2], [21].

As part of our ablation study, we also evaluated how different backbones affect the COVID-19 diagnosis accuracy of *PrepNet*. More precisely, we replicate the experiments for each dataset (SARS-COV-2 and UCSD COVID-CT) and evaluate different CNN architectures as part of our COVID-classifier Module (See Section III-A for more information). The CNN architectures include ResNet18 [17], Inception [35], and EfficientNet-B0 [36]. We report results in Table III. Experimental results show that in almost all backbones, the average cross-dataset performance increases with the cost of a small decrease in the within-dataset accuracy.

Finally, in order to evaluate the generalisation capabilities of *PrepNet* and our baselines, we evaluate how our trained models perform on an unseen dataset, i.e. the *MosMed dataset* [13]. The results in Table IV show the improvements of our *AutoEncoder* and *PrepNet* models in terms of BA and sensitivity, however, with a decrease in specificity and AUC when compared with the COVID-19 classifier. Despite the decrease in specificity, we argue that especially for medical diagnosis and screening, a low specificity is less harmful than a reduction in sensitivity, as false positive cases can be discarded by additional examinations. On the contrary, a higher sensitivity is important as false negatives should be low.

[1] https://stanfordmlgroup.github.io/competitions/mura/

				Dataset portions		
Dataset	Type	Size	Country	Train	Validation	Test
SARS-COV-2 [11]	2D CT	Various	Brazil	$2,046$ (70%)	439 (15%)	439 (15%)
UCSD COVID-CT [12]	2D CT	Various	China	423 (57%)	116 (16%)	201 (27%)
MosMed Dataset [13]	3D CT	Various	Russia	1100 images for unseen test dataset		

TABLE I
Public datasets used in our study together with their corresponding data splits. The SARS-COV-2 [11] and the UCSD COVID-CT [12] datasets are used for training and evaluating our models, while the MosMed dataset [13] is used for evaluation purposes only.

Test dataset →		SARS-COV-2				UCSD COVID-CT			Within Test	Cross-Dataset	Pre-trained
Dataset portion	BA	Sens	Spec	AUC	Test	Sens	Spec	AUC	Average	Average	encoder
	COVID classifier										
SARS-COV-2	0.8924	0.9292	0.7876	0.8584	0.4433	0.7835	**0.1262**	0.4548	**0.8587**	0.4159	Yes
UCSD COVID-CT	0.3295	0.3476	0.2743	0.3110	**0.8250**	0.7113	**0.9320**	**0.8216**	(baseline)	(baseline)	
	AutoEncoder										
SARS-COV-2	0.8956	**0.9907**	0.6460	0.8183	0.4983	0.9175	0.0970	0.5073	0.8555	0.4836	Yes
UCSD COVID-CT	0.49405	0.6030	**0.3008**	0.4519	0.8154	0.7216	0.8846	0.8031	(−0.32%)	(+6.77%)	
	PrepNet										
SARS-COV-2	**0.9007**	0.9353	**0.7982**	**0.8668**	**0.5157**	**0.9175**	0.1067	**0.5121**	0.8404	**0.5343**	Yes
UCSD COVID-CT	**0.5545**	**0.6446**	0.1858	**0.4852**	0.7800	**0.8556**	0.7087	0.7822	(−1.83%)	(+11.84%)	

TABLE II
Test performance of different baselines compared to our *PrepNet* model. Results demonstrate that our model is capable of increasing the cross-dataset average.

D. Discussion

The baseline and proposed pre-processing approaches introduce performance drops when applied before within-dataset classification. These approaches usually reduce the test accuracies when trained and evaluated on the same dataset using the corresponding dataset splits. Therefore, we further investigate the intermediate results of the baseline auto-encoder and *PrepNet* on a case-by-case basis. Severe cases of generated artifacts through reconstruction via the baseline auto-encoder and the *PrepNet* are presented in Figure 3. We conjecture that the drop in within-dataset test performance is caused by occasional artifacts such as these. These quality drops can be clearly seen in the reconstruction loss. However, it is not straightforward to correct them. We could eventually overcome this by also investigating different data-augmentation strategies and by improving the network architecture of our auto-encoder. Additionally, we depict sample images in which the models failed to make a correct decision after auto-encoder or *PrepNet* (See Fig. 4). Limited amount of training data and noisy labels of public datasets are other factors contributing to low classification accuracies. One possible way to tackle this limitation is to rely on weakly supervised learning methods to improve the COVID-19 classification accuracy with the methodology summarized in [37].

V. Conclusions and Future Work

In this paper, we introduced a novel approach to unify several CT scan datasets with respect to varying image datasets and acquisition circumstances such as CT scanner technology through training an adaptive pre-processing network that removes such specificities from the images themselves. Additionally, we presented initial results demonstrating the applicability of the method on three publicly available benchmark datasets. This way, it is possible to shift the focus of model training from merely optimizing hold-out test set performance on *the same* data distribution (which likely does not transfer to any other environment) towards cross-dataset detection accuracy. The proposed *PrepNet* improves the cross-dataset balanced accuracy by a margin of 11.84 percentage points (*SARS-CoV-2 CT-scan dataset* [11]) at the expanse of a decline in the within dataset test performance of ca. 1.83pp (*UCSD COVID-CT database* [12]). These results suggest that the trainable preprocessing network erases some of the necessary information for diagnosis, due to artifacts. This information could be partially retained by propagating the gradients of the COVID-19 classifier network through the preprocessing model, and generated artifacts could be detected automatically by monitoring the reconstruction loss of the auto-encoder module. This, together with further investigations on the applicability and generality of the proposed approach to combine multiple datasets, is an intriguing theme for future research.

Acknowledgment

This research was financially supported by the ZHAW Digital Futures Fund under contracts "*SDMCT—Standardized*

Test dataset →	SARS-COV-2				UCSD COVID-CT				Within Test	Cross-Dataset	Pre-trained
Dataset portion	BA	Sens	Spec	AUC	Test	Sens	Spec	AUC	Average	Average	encoder
	VGG19										
SARS-COV-2	**0.9007**	**0.9353**	0.7982	**0.8668**	0.5157	0.9175	0.1067	0.5121	**0.8404**	**0.5343**	Yes
UCSD COVID-CT	**0.5545**	**0.6446**	0.1858	**0.4852**	0.7800	0.8556	0.7087	0.7822	(−1.83%)	(+11.84%)	
	ResNet18										
SARS-COV-2	0.7462	0.7046	**0.8584**	0.7815	0.4728	0.8144	0.1538	0.4841	0.7345	0.4940	Yes
UCSD COVID-CT	0.5152	0.6246	0.1947	0.4096	0.7228	0.8351	0.6154	0.7252	(−12.42%)	(+7.81%)	
	Inception										
SARS-COV-2	0.8553	0.9046	0.7080	0.8063	0.4703	**0.9485**	0.02885	0.4886	0.8286	0.3995	Yes
UCSD COVID-CT	0.3288	0.36308	0.2212	0.2922	0.8020	0.8351	0.7692	0.8021	(−3.01%)	(−1.64%)	
	EfficientNet-B0										
SARS-COV-2	0.8735	0.8923	0.8142	0.8532	0.5223	0.5979	0.4519	0.5249	0.8253	0.4835	Yes
UCSD COVID-CT	0.4447	0.5015	**0.2743**	0.3879	0.7772	0.8041	0.7500	0.7771	(−3.34%)	(+6.76%)	

TABLE III
EXPERIMENTAL RESULTS OF *PrepNet* WITH DIFFERENT BACKBONES: VGG19 [16], RESNET18 [17], INCEPTION [35], AND EFFICIENTNET-B0 [36]. NOTE THAT *PrepNet* INCREASES THE CROSS-DATASET AVERAGE.

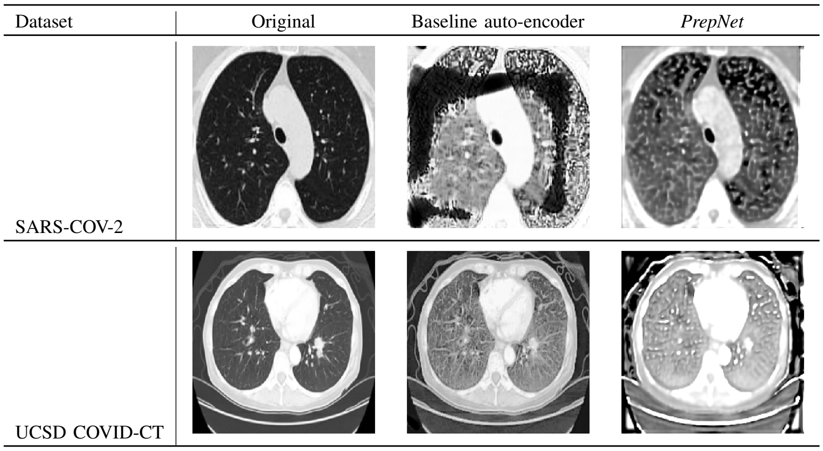

Fig. 3. Severe cases of artifacts generated by the baseline and the proposed *PrepNet*. The images demonstrate different levels of distortions like e.g. extreme contrasts.

Test dataset →	MosMed				Pre-trained
Preprocessing	BA	Sens	Spec	AUC	encoder
COVID-classifier	0.6066 (baseline)	0.5246 (baseline)	**0.8771** (baseline)	**0.7009** (baseline)	Yes
AutoEncoder	0.6693 (+6.27%)	0.7142 (+18.96%)	0.5175 (−35.96%)	0.6159 (−8.50%)	Yes
PrepNet	**0.7073** (+10.07%)	**0.7558** (+23.12%)	0.5438 (−33.33%)	0.6498 (−5.11%)	Yes

TABLE IV
EXPERIMENTAL COVID-19 CLASSIFICATION RESULTS OF THE TRAINED COVID-19 CLASSIFIER, AUTO-ENCODER, AND *PrepNet* MODELS ON THE MOSMED [13] UNSEEN DATASET.

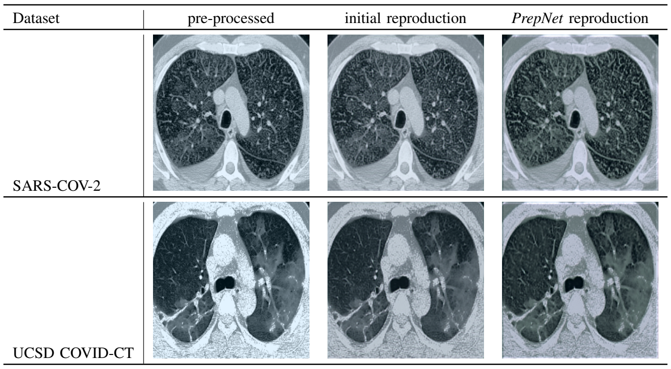

Fig. 4. Samples CT scans that are wrongly classified after the trainable preprocessing.

Data and Modeling for AI-based CoVID-19 Diagnosis Support on CT Scans” as well as “*Synthetic data generation of CoVID-19 CT/X-rays images for enabling fast triage of healthy vs. unhealthy patients*”.

REFERENCES

[1] WHO. (2021) WHO COVID-19 situation reports. [Online]. Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports

[2] M. J. Horry, S. Chakraborty, M. Paul, A. Ulhaq, B. Pradhan, M. Saha, and N. Shukla, "Covid-19 detection through transfer learning using multimodal imaging data," *IEEE Access*, vol. 8, pp. 149 808–149 824, 2020.

[3] C. Chen, G. Gao, Y. Xu, L. Pu, Q. Wang, L. Wang, W. Wang, Y. Song, M. Chen, L. Wang *et al.*, "Sars-cov-2–positive sputum and feces after conversion of pharyngeal samples in patients with covid-19," *Annals of internal medicine*, vol. 172, no. 12, pp. 832–834, 2020.

[4] X. Mei, H.-C. Lee, K.-y. Diao, M. Huang, B. Lin, C. Liu, Z. Xie, Y. Ma, P. M. Robson, M. Chung, A. Bernheim, V. Mani, C. Calcagno, K. Li, S. Li, H. Shan, J. Lv, T. Zhao, J. Xia, Q. Long, S. Steinberger, A. Jacobi, T. Deyer, M. Luksza, F. Liu, B. P. Little, Z. A. Fayad, and Y. Yang, "Artificial intelligence–enabled rapid diagnosis of patients with covid-19," *Nature Medicine*, vol. 26, pp. 1224–1228, 2020.

[5] S. A. Harmon, T. H. Sanford, S. Xu, E. B. Turkbey, H. Roth, Z. Xu, D. Yang, A. Myronenko, V. Anderson, A. Amalou, M. Blain, M. Kassin, D. Long, N. Varble, S. M. Walker, U. Bagci, A. M. Ierardi, E. Stellato, G. G. Plensich, G. Franceschelli, C. Girlando, G. Irmici, D. Labella, D. Hammoud, A. Malayeri, E. Jones, R. M. Summers, P. L. Choyke, D. Xu, M. Flores, K. Tamura, H. Obinata, H. Mori, F. Patella, M. Cariati, G. Carrafiello, P. An, B. J. Wood, and B. Turkbey, "Artificial intelligence for the detection of covid-19 pneumonia on chest ct using multinational datasets," *Nature Communications*, vol. 11, p. 4080, 2020.

[6] L. Li, L. Qin, Z. Xu, Y. Yin, X. Wang, B. Kong, J. Bai, Y. Lu, Z. Fang, Q. Song, K. Cao, D. Liu, G. Wang, Q. Xu, X. Fang, S. Zhang, J. Xia, and J. Xia, "Using artificial intelligence to detect covid-19 and community-acquired pneumonia based on pulmonary ct: Evaluation of the diagnostic accuracy," *Radiology*, vol. 296, no. 2, pp. E65–E71, 2020.

[7] X. Wang, X. Deng, Q. Fu, Q. Zhou, J. Feng, H. Ma, W. Liu, and C. Zheng, "A weakly-supervised framework for covid-19 classification and lesion localization from chest ct," *IEEE Transactions on Medical Imaging*, vol. 39, no. 8, pp. 2615–2625, 2020.

[8] N. Lessmann, C. I. Sánchez, L. Beenen, L. H. Boulogne, M. Brink, E. Calli, J.-P. Charbonnier, T. Dofferhoff, W. M. van Everdingen, P. K. Gerke *et al.*, "Automated assessment of covid-19 reporting and data system and chest ct severity scores in patients suspected of having covid-19 using artificial intelligence," *Radiology*, vol. 298, no. 1, pp. E18–E28, 2021.

[9] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds., 2014, pp. 2672–2680. [Online]. Available: https://proceedings.neurips.cc/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html

[10] J. Schmidhuber, "Generative adversarial networks are special cases of artificial curiosity (1990) and also closely related to predictability minimization (1991)," *Neural Networks*, vol. 127, pp. 58–66, 2020.

[11] E. Soares and P. Angelov, "A large dataset of real patients CT scans for COVID-19 identification," 2020. [Online]. Available: https://doi.org/10.7910/DVN/SZDUQX

[12] J. Zhao, Y. Zhang, X. He, and P. Xie, "Covid-ct-dataset: a ct scan dataset about covid-19," *arXiv preprint arXiv:2003.13865*, 2020.

[13] S. P. Morozov, A. E. Andreychenko, N. A. Pavlov, A. V. Vladzymyrskyy, N. V. Ledikhova, V. A. Gombolevskiy, I. A. Blokhin, P. B. Gelezhe, A. V. Gonchar, and V. Y. Chernina, "Mosmeddata: Chest ct scans with covid-19 related findings dataset," 2020.

[14] J. P. Cohen, P. Morrison, and L. Dao, "Covid-19 image data collection," *arXiv 2003.11597*, 2020. [Online]. Available: https://github.com/ieee8023/covid-chestxray-dataset

[15] H. Gunraj, A. Sabri, D. Koff, and A. Wong, "Covid-net ct-2: Enhanced deep neural networks for detection of covid-19 from chest ct images through bigger, more diverse learning," *arXiv preprint arXiv:2101.07433*, 2021.

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

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[18] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[19] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[20] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017.

[21] X. He, X. Yang, S. Zhang, J. Zhao, Y. Zhang, E. Xing, and P. Xie, "Sample-efficient deep learning for covid-19 diagnosis based on ct scans," *MedRxiv*, 2020.

[22] A. Mobiny, P. A. Cicalese, S. Zare, P. Yuan, M. Abavisani, C. C. Wu, J. Ahuja, P. M. de Groot, and H. Van Nguyen, "Radiologist-level covid-19 detection using ct scans with detail-oriented capsule networks," *arXiv preprint arXiv:2004.07407*, 2020.

[23] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," *arXiv preprint arXiv:1710.09829*, 2017.

[24] H. Gao, L. Zhuang, L. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *CVPR*, vol. 1, 2017, p. 3.

[25] T. D. Pham, "A comprehensive study on classification of covid-19 on computed tomography with pretrained convolutional neural networks," *Scientific Reports*, vol. 10, no. 1, pp. 1–8, 2020.

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[27] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8697–8710.

[28] S. Chaganti, A. Balachandran, G. Chabin, S. Cohen, T. Flohr, B. Georgescu, P. Grenier, S. Grbic, S. Liu, F. Mellot *et al.*, "Quantification of tomographic patterns associated with covid-19 from chest ct," *arXiv preprint arXiv:2004.01279*, 2020.

[29] A. Wong, "Netscore: towards universal metrics for large-scale performance analysis of deep neural networks for practical on-device edge usage," in *International Conference on Image Analysis and Recognition*. Springer, 2019, pp. 15–26.

[30] A. Amyar, R. Modzelewski, H. Li, and S. Ruan, "Multi-task deep learning based ct imaging analysis for covid-19 pneumonia: Classification and segmentation," *Computers in Biology and Medicine*, vol. 126, p. 104037, 2020.

[31] C. de Vente, L. H. Boulogne, K. V. Venkadesh, C. Sital, N. Lessmann, C. Jacobs, C. I. Sánchez, and B. van Ginneken, "Improving automated covid-19 grading with convolutional neural networks in computed tomography scans: An ablation study," *arXiv preprint arXiv:2009.09725*, 2020.

[32] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The balanced accuracy and its posterior distribution," in *2010 20th international conference on pattern recognition*. IEEE, 2010, pp. 3121–3124.

[33] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[34] L. Tuggener, M. Amirian, K. Rombach, S. Lörwald, A. Varlet, C. Westermann, and T. Stadelmann, "Automated machine learning in practice: state of the art and recent results," in *2019 6th Swiss Conference on Data Science (SDS)*. IEEE, 2019, pp. 31–36.

[35] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Computer Vision and Pattern Recognition (CVPR)*, 2015. [Online]. Available: http://arxiv.org/abs/1409.4842

[36] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning*. PMLR, 2019, pp. 6105–6114.

[37] N. Simmler, P. Sager, P. Andermatt, R. Chavarriaga, F.-P. Schilling, M. Rosenthal, and T. Stadelmann, "A survey of un-, weakly-, and semi-supervised learning methods for noisy, missing and partial labels in industrial vision applications," in *Proceedings of the 8th SDS*. IEEE, 2021.