Evolving Systems manuscript No.
(will be inserted by the editor)

A large multiclass dataset of CT scans for COVID-19 identification

Anonymous

Received: date / Accepted: date

Abstract The infection by SARS-CoV-2 which causes the COVID-19 disease has spread widely over the whole world since the beginning of 2020. Following the epidemic which started in Wuhan, China on January 30, 2020 the World Health Organization (WHO) declared a global health emergency and a pandemic. In this paper, we describe a publicly available multiclass CT scan dataset for SARS-CoV-2 infection identification. Which currently contains 4173 CT-scans of 210 different patients, out of which 2168 correspond to 80 patients infected with SARS-CoV-2 and confirmed by RT-PCR. These data have been collected in the Public Hospital of the Government Employees of Sao Paulo and the Metropolitan Hospital of Lapa, both in Sao Paulo – Brazil. The aim of this data set is to encourage the research and development of artificial intelligent methods that are able to identify SARS-CoV-2 or other diseases through the analysis of CT scans. As a baseline result for this data set, we used the recently introduced eXplainable Deep Learning approach (xDNN), which is a transparent deep learning approach that allows users to inspect the decisions of the network.

Keywords CT-scans · COVID-19 detection · Machine Learning · Explainable AI

1 Introduction

Coronavirus disease 2019 (COVID-19) is an infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The disease was first identified in December 2019 in Wuhan, the capital of China's Hubei province, and has since spread worldwide [2]. On January 30, 2020 the World Health Organization (WHO) declared a global health emergency [1]. Common symptoms of COVID-19 include fever, cough, and shortness of breath [10, 23].

Anonymous

While the majority of cases result in mild symptoms, some progress to viral pneumonia. By 7 August 2020, over 19 million officially confirmed cases were reported in practically every corner of the Earth with 717,687 officially reported deaths documented [9].

As the first countries explore deconfinement strategies [8,16] after a long period of quarantine, the death toll of COVID-19 keeps rising, specially in US, UK, and Brazil [9]. New technologies and strategies have emerged in order to support healthcare systems during this pandemic time [11,22]. As early as March 2020, Chinese hospitals used artificial intelligence (AI)-assisted computed tomography (CT) imaging analysis to screen COVID-19 cases and streamline diagnosis [12].

In this work, we build a large multiclass dataset of CT scans for SARS-CoV-2 infection identification. The dataset is built upon on the recently introduced dataset [19]. The proposed dataset contains 4173 CT-scans of 210 different patients which are divided into 3 different classes (healthy, COVID-19, and other pulmonary diseases). These data have been collected from real patients in hospitals from Sao Paulo, Brazil. The aim of this dataset is to encourage the research and development of artificial intelligence (AI) methods that are able to identify if a person is infected by SARS-CoV-2 through the analysis of his/her CT scans.

An open-source dataset for COVID-19 identification through CT scans has been proposed by [24], however, the data collected for this dataset has been acquired from scientific journals and may not provide the necessary quality to train an algorithm for complex applications as such. Moreover, other authors as [17,13,15,18] provided open-source datasets and solutions based on X-ray scans which are not detailed as CT scans.

As a baseline result for the new dataset based on CT scans, we consider the eXplainable Deep Learning approach (xDNN) [4]. As the explanation of AI systems is essential to medical applications, we used the xDNN approach as baseline for this application. XDNN is a prototype-based approach that allows users to audit the decisions of the network through its similarity mechanism. XDNN obtained an $F1$ score of 97.31%, which is higher than traditional deep learning approaches such as ResNet.

2 Methods

The proposed dataset is composed of 4173 CT-scans of 210 different patients which are divided into: 80 patients infected by SARS-CoV-2; 80 patients with other pulmonary diseases as non-COVID pneumonia, bronchitis, and lung cancer; and 50 patients with healthy lung conditions. The data was collected from March 15 to June 15 2020 in the Public Hospital of the Government Employees of Sao Paulo, and the Metropolitan Hospital of Lapa, Sao Paulo – Brazil. The following demographic data have been collected during the clinical analysis of each patient:

– Sex

- Age
- Number of days since the 1st symptoms
- Comorbidities
- Hypertension
- Diabetes
- Chronic obstructive pulmonary disease (COPD)
- Obesity
- Pulmonary involvement > 50%
- Outcome

Table 1 details the patient's considered in this study.

Condition	Patients	CT-Scans	Average CT-Scans per patient
Healthy	50	758	15
COVID-19	80	2168	27
Other pulmonary diseases	80	1247	16
TOTAL	210	4173	20

Table 1 This table demonstrates the number of patients considered to compose the dataset. In this case, we considered data from 80 patients infected by SARS-CoV-2, out of which 41 were male and 39 were female. We also considered data from 80 patients presenting other pulmonary diseases such as lung cancer, bronchitis, etc. The dataset is also composed of CT scans that do not present any pulmonary disease, These data refer to data of 50 patients.

The inclusion criteria for this study are listed as follows:

- Patients with a positive new coronavirus nucleic acid antibody and confirmed by the CDC;
- Patients who underwent thin-section CT;
- Age>= 18;
- Presence of lung infection in CT images.

The median duration from the onset of the illness to CT scan was 5 days, ranging from 1 to 14 days. The CT protocol was as follows: 120 kV; automatic tube current (180 mA-400 mA); iterative reconstruction; 64 mm detector; rotation time, 0.35 sec; slice thickness, 5 mm; collimation, 0.625 mm; pitch, 1.5; matrix, 512×512; and breath hold at full inspiration. The reconstruction kernel used is set as "lung smooth with a thickness of 1 mm and an interval of 0.8 mm". During reading, the lung window (with window wiDecision Treeh 1200 HU and window level-600 HU) was used. Figure (2) illustrates some examples of CT scans found in the dataset.

3 Data Records

The database can be downloaded from Synapse(`https://www.synapse.org/#!Synapse:syn22174850`), and it has been presented in two formats: PNG and

CSV, where PNG represents the CT scans files and CVS are the demographic data. Fig. (1) illustrates the data distribution for the patients infected by SARS-CoV-2 and considered in this study.

Fig. (1) illustrates the data distribution for the patients infected by SARS-CoV-2 and considered in this study.

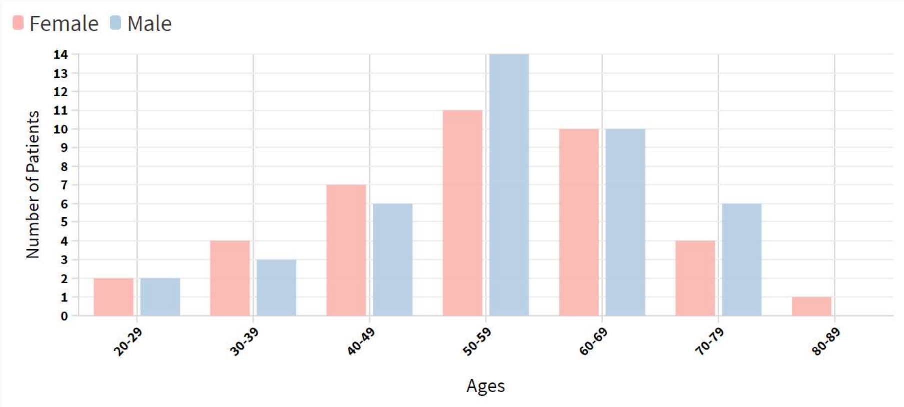

Fig. 1 The study considered data for 80 different patients (41 male and 39 female patients). The data revealed that the major of the patients are 50-59 years old.

The data types of the demographic data variables considered in this study are depicted below:

- Sex (Boolean)
- Age (Integer)
- Number of days since the 1st symptoms (Integer)
- Comorbidities (Boolean)
- Hypertension (Boolean)
- Diabetes (Boolean)
- Chronic obstructive pulmonary disease (COPD) (Boolean)
- Obesity (Boolean)
- Pulmonary involvement > 50% (Boolean)
- Outcome (Boolean)

Fig. (2) illustrates different examples of data available in the proposed dataset.

4 Technical Validation

In order to validate our data in this section we report the results by different classification approaches. The following metrics have been used to evaluate the classification of the CT scans:

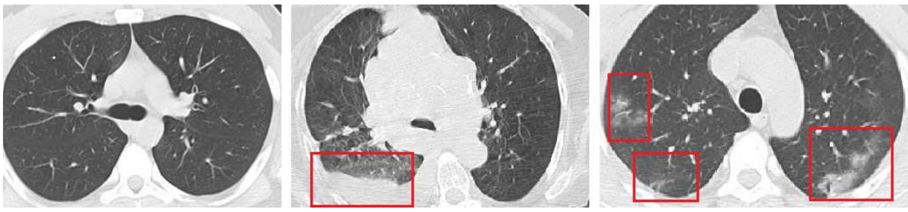

Fig. 2 **a)** A 27-year-old male patient presented with fever and headache for 2 days. CT scans do not show the presence of any pulmonary disease. The RT-PCR test revealed negative for SARS-CoV-2 infection. **b)** A 63-year-old woman patient presented shortness of breath and cough for 4 days. CT scan shows the presence of subpulmonic pleural effusion. The RT-PCR test revealed negative for SARS-CoV-2. **c)** A 31-year-old woman presented fever, dry cough, shortness of breath for 4 days. CT scan revealed multifocal bilateral consolidation with ground-glass opacities with typical distribution. The RT-PCR tested positve for SARS-CoV-2.

$$Accuracy(\%) = \frac{TP + TN}{TP + FP + TN + FN} \times 100, \quad (1)$$

Precision:

$$Precision(\%) = \frac{TP}{TP + FP} \times 100, \quad (2)$$

Recall:

$$Recall(\%) = \frac{TP}{TP + FN} \times 100, \quad (3)$$

$F1$ Score:

$$F1\ Score(\%) = 2 \times \frac{Precision \times Recall}{Precision + Recall} \times 100, \quad (4)$$

where TP, FP, TN, FN denote true and false, negative and positive respectively.

The area under the curve, AUC, is defined through the TP rate and FN rate.

In this section we report the results obtained by the xDNN classification approach [4, 21] when applied to the proposed SARS-CoV-2 CT scan data set. We divided the dataset into 80% for training purposes and 20% for validation purposes. The division has been made in terms of patients; therefore, we separated data of 168 patients for training and data for 42 patients for validation. Results presented in Table 2 compare the performance of the xDNN algorithm with other state-of-the-art approaches, including ResNet, GoogleNet, VGG-16, AlexNet, Decision Tree, and AdaBoost.

The xDNN [4, 3] classifier provided better results in terms of all metrics than the other state-of-the-art approaches, including ResNet, GoogleNet,

MethodMetric	Accuracy	Precision	Recall	Specificity	F1 Score	AUC
xDNN	97.38%	99.16%	95.53%	96.42%	97.31%	97.36%
ResNet	94.96%	93.00%	97.15%	94.36%	95.03%	94.98%
GoogleNet	91.73%	90.20%	93.50%	90.17%	91.82%	91.79%
VGG-16	94.96%	94.02%	95.43%	94.51%	94.97%	94.96%
AlexNet	93.75%	94.98%	92.28%	92.32%	93.61%	93.68%
Decision Tree	79.44%	76.81%	83.13%	77.16%	79.84%	79.51%
AdaBoost	95.16%	93.63%	96.71%	94.98%	95.14%	95.19%

Table 2 Results considering different methods for the COVID-19 identification.

VGG-16, and Alexnet. Moreover, it also provided highly interpretable results [6] that may be helpful for specialists (medical doctors). Rules generated by the identified prototypes are illustrated by Figs. (3) and (4), respectively. xDNN identified data of 18 patients with COVID-19 as prototypes and data of 11 patients non-infected as prototypes. The training time for the xDNN algorithm [4] was only 11.82 seconds for all images (an average of 5 milliseconds per image. On the other hand, the traditional deep learning approach may take hours for the same task and usually requires hardware accelerators such as GPUs and once trained is not flexible to new data. We have to stress that xDNN does not require full re-training if new data is presented [5] - it keeps all prototypes identified so far and may add new ones if the data pattern requires that [19,20].

Balanced one-way ANalysis Of VAriance (ANOVA) [14] was used to compare the results provided by the classification methods. The null hypothesis is that the mean results provided by the methods are the same. A cutoff value p less than 0.05 suggests that the accuracy of at least one of the algorithms is significantly different from the others. A $p = 4.38e - 22$ was obtained and, therefore, the mean accuracy of the algorithms is not all the same; the null hypothesis was rejected.

The Tukey Honestly Significant Difference (HSD) test [14] was performed to compare pairs of classifiers over accuracy. Table 3 shows the results of the Tuckey HSD test for a 95% confidence interval for the true difference of the means.

Method 1	Method 2	meandiff	p-adj	lower	upper	Reject
xDNN	Resnet	-2.28	0.068	-4.6604	0.1004	False
xDNN	GoogleNet	-5.6583	0.001	-8.0387	-3.278	True
xDNN	Vgg16	-2.385	0.0493	-4.7654	-0.0046	True
xDNN	Alexnet	-3.7567	0.001	-6.137	-1.3763	True
xDNN	Decision Tree	-17.8783	0.001	-20.2587	-15.498	True
xDNN	Adaboost	-2.0583	0.1272	-4.4387	0.322	False
Resnet	GoogleNet	-3.3783	0.0015	-5.7587	-0.998	True
Resnet	Vgg16	-0.105	0.9	-2.4854	2.2754	False
Resnet	Alexnet	-1.4767	0.4709	-3.857	0.9037	False
Resnet	Decision Tree	-15.5983	0.001	-17.9787	-13.218	True
Resnet	Adaboost	0.2217	0.9	-2.1587	2.602	False
GoogleNet	Vgg16	3.2733	0.0023	0.893	5.6537	True
GoogleNet	Alexnet	1.9017	0.1912	-0.4787	4.282	False
GoogleNet	Decision Tree	-12.22	0.001	-14.6004	-9.8396	True
GoogleNet	Adaboost	3.6	0.001	1.2196	5.9804	True
Vgg16	Alexnet	-1.3717	0.5491	-3.752	1.0087	False
Vgg16	Decision Tree	-15.4933	0.001	-17.8737	-13.113	True
Vgg16	Adaboost	0.3267	0.9	-2.0537	2.707	False
Alexnet	Decision Tree	-14.1217	0.001	-16.502	-11.7413	True
Alexnet	Adaboost	1.6983	0.3061	-0.682	4.0787	False
Decision Tree	Adaboost	15.82	0.001	13.4396	18.2004	True

Table 3 Tukey Test Results.

If the $p - adj < 0.05$ then the null hypothesis is rejected and the difference between the methods is statistically significant. As shown in Table 3 the proposed xDNN has results statistically different from 4 traditional approaches, including well-known deep learning approaches as GoogleNet, VGG-16, and AlexNet.

Through the xDNN method we generated (extracted from the data) linguistic *IF…THEN* rules which involve actual images of both cases (COVID-19 and NO COVID-19) as illustrated in Figs. (3) and (4). Such transparent rules can be used in a clear decision-making process for early diagnostics for COVID-19 infection. Rapid detection with high sensitivity of viral infection may allow better control of the viral spread. Early diagnosis of COVID-19 is crucial for disease treatment and control.

5 Conclusion

In the context of a pandemic and the urgency to contain the crisis, research has increased exponentially in order to alleviate the healthcare systems burden [7]. However, many prediction models for diagnosis and prognosis of COVID-19 infection are at high risk of bias and model overfitting as well as poorly reported, their alleged performance being likely optimistic. In order to prevent premature implementation in hospitals, tools must be robustly evaluated along several practical tests. Indeed, while some AI-assisted tools might be powerful, they do not replace clinical judgment and their diagnostic performance cannot be assessed or claimed without a proper clinical trial.

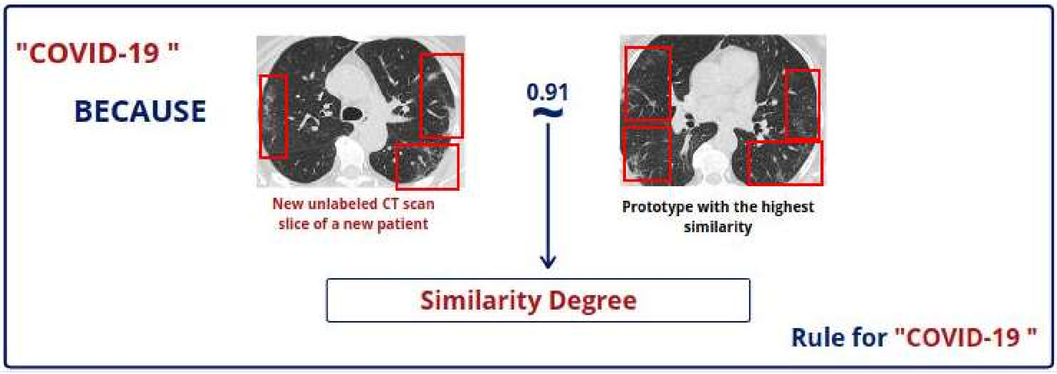

Fig. 3 Final rule given by xDNN classifier for the COVID-19 identification. Differently, from typical deep neural networks, xDNN provides highly interpretable rules which can be visualised and used by human experts for the early evaluation of patients suspected of COVID-19 infection. The classification is done based on the similarity of the unlabeled CT scan slice to the identified prototypes.

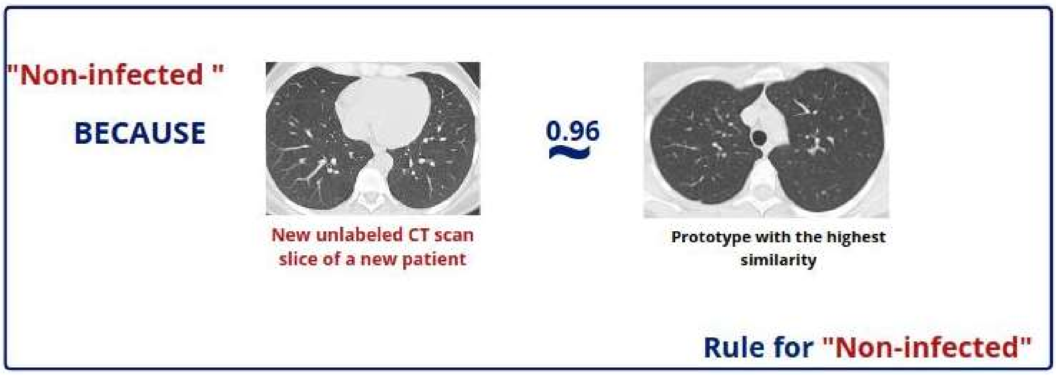

Fig. 4 Non-SARS-CoV-2 final rule given by the proposed eXplainable Deep Learning classifier.

Moreover, The lack of a public database made it difficult to conduct large-scale robust evaluations. This small number of samples prevents proper cohort selection which is a limitation of this study and exposes our evaluation to sample bias. In this study, we present a database which is composed of 4173 CT-scans of 210 different patients, out of which 2168 correspond to 80 patients infected with SARS-CoV-2 and confirmed by RT-PCR. These data have been collected at the Public Hospital of the Government Employees of Sao Paulo and the Metropolitan Hospital of Lapa, Sao Paulo, Brazil. Sao Paulo is now one of the global epicenters of the COVID-19 disease.

As a baseline result for the proposed dataset, we used an explainable deep learning approach. The xDNN classifier presented an $F1$ score of 97.31% for the proposed task. Moreover, the xDNN approach provided insights into the decision-making process which is helpful to support specialists in the diagnosis of the disease. This is of great importance for medical specialists to understand and diagnose COVID-19 at its early stages via computer tomography. The proposed dataset is available https://www.synapse.org/#!Synapse:syn22174850 and xDNN [4] code is available at https://github.com/Plamen-Eduardo/xDNN-SARS-CoV-2-CT-Scan.

Code availability

We provided the code used in this research at https://github.com/Plamen-Eduardo/xDNN-SARS-CoV-2-CT-Scan. Other codes are available upon request to the corresponding author.

Data availability

The data that support the findings of this study are openly available in Synapse at https://www.synapse.org/#!Synapse:syn22174850 and a small version of it in Kaggle at https://www.kaggle.com/datasets/plameneduardo/sarscov2-ctscan-dataset.

Competing Interests

The authors declare no competing interests.

References

1. World health organization declares global emergency: A review of the 2019 novel coronavirus (covid-19). International Journal of Surgery **76**, 71–76 (2020). DOI https://doi.org/10.1016/j.ijsu.2020.02.034
2. Andersen, K.G., Rambaut, A., Lipkin, W.I., Holmes, E.C., Garry, R.F.: The proximal origin of sars-cov-2. Nature medicine **26**(4), 450–452 (2020)
3. Angelov, P., Soares, E.: Towards deep machine reasoning: a prototype-based deep neural network with decision tree inference. In: 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 2092–2099 (2020). DOI 10.1109/SMC42975.2020.9282812
4. Angelov, P., Soares, E.: Towards explainable deep neural networks (xDNN). Neural Networks **130**, 185–194 (2020)
5. Angelov, P., Soares, E.: Detecting and learning from unknown by extremely weak supervision: exploratory classifier (xclass). Neural Computing and Applications **33**(22), 15145–15157 (2021)
6. Angelov, P.P., Soares, E.A., Jiang, R., Arnold, N.I., Atkinson, P.M.: Explainable artificial intelligence: an analytical review. WIREs Data Mining and Knowledge Discovery **11**(5), e1424 (2021). DOI https://doi.org/10.1002/widm.1424
7. Cohen, J.P., Dao, L., Roth, K., Morrison, P., Bengio, Y., Abbasi, A.F., Shen, B., Mahsa, H.K., Ghassemi, M., Li, H., et al.: Predicting covid-19 pneumonia severity on chest x-ray with deep learning. Cureus **12**(7) (2020)
8. Cousins, S.: New zealand eliminates covid-19. The Lancet **395**(10235), 1474 (2020)
9. Dong, E., Du, H., Gardner, L.: An interactive web-based dashboard to track covid-19 in real time. The Lancet infectious diseases (2020)

10. Guan, W.j., Ni, Z.y., Hu, Y., Liang, W.h., Ou, C.q., He, J.x., Liu, L., Shan, H., Lei, C.l., Hui, D.S., et al.: Clinical characteristics of coronavirus disease 2019 in china. New England journal of medicine **382**(18), 1708–1720 (2020)
11. Hu, Z., Ge, Q., Li, S., Jin, L., Xiong, M.: Artificial intelligence forecasting of covid-19 in china. arXiv preprint arXiv:2002.07112 **150**(01-20) (2020)
12. Jin, Y.H., Cai, L., Cheng, Z.S., Cheng, H., Deng, T., Fan, Y.P., Fang, C., Huang, D., Huang, L.Q., Huang, Q., et al.: A rapid advice guideline for the diagnosis and treatment of 2019 novel coronavirus (2019-ncov) infected pneumonia (standard version). Military Medical Research **7**(1), 4 (2020)
13. Mangal, A., Kalia, S., Rajgopal, H., Rangarajan, K., Namboodiri, V., Banerjee, S., Arora, C.: Covidaid: Covid-19 detection using chest x-ray. arXiv preprint arXiv:2004.09803 (2020)
14. McHugh, M.L.: Multiple comparison analysis testing in anova. Biochemia medica: Biochemia medica **21**(3), 203–209 (2011)
15. Pham, T.D.: Classification of covid-19 chest x-rays with deep learning: new models or fine tuning? Health Information Science and Systems **9**, 1–11 (2021)
16. Salathé, M., Althaus, C.L., Neher, R., Stringhini, S., Hodcroft, E., Fellay, J., Zwahlen, M., Senti, G., Battegay, M., Wilder-Smith, A., et al.: Covid-19 epidemic in switzerland: on the importance of testing, contact tracing and isolation. Swiss medical weekly **150**(11-12), w20225 (2020)
17. Santa Cruz, B.G., Bossa, M.N., Sölter, J., Husch, A.D.: Public covid-19 x-ray datasets and their impact on model bias–a systematic review of a significant problem. Medical image analysis **74**, 102225 (2021)
18. Signoroni, A., Savardi, M., Benini, S., Adami, N., Leonardi, R., Gibellini, P., Vaccher, F., Ravanelli, M., Borghesi, A., Maroldi, R., et al.: Bs-net: Learning covid-19 pneumonia severity on a large chest x-ray dataset. Medical Image Analysis **71**, 102046 (2021)
19. Soares, E., Angelov, P., Biaso, S., Froes, M.H., Abe, D.K.: Sars-cov-2 ct-scan dataset: A large dataset of real patients ct scans for sars-cov-2 identification. medRxiv **1**(1) (2020)
20. Soares, E., Angelov, P., Costa, B., Castro, M.: Actively semi-supervised deep rule-based classifier applied to adverse driving scenarios. In: 2019 International Joint Conference on Neural Networks (IJCNN), pp. 1–8 (2019). DOI 10.1109/IJCNN.2019.8851842
21. Soares, E., Angelov, P., Filev, D., Costa, B., Castro, M., Nageshrao, S.: Explainable density-based approach for self-driving actions classification. In: 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA), pp. 469–474 (2019). DOI 10.1109/ICMLA.2019.00087
22. Ting, D.S.W., Carin, L., Dzau, V., Wong, T.Y.: Digital technology and covid-19. Nature medicine **26**(4), 459–461 (2020)
23. Xu, Z., Shi, L., Wang, Y., Zhang, J., Huang, L., Zhang, C., Liu, S., Zhao, P., Liu, H., Zhu, L., et al.: Pathological findings of covid-19 associated with acute respiratory distress syndrome. The Lancet respiratory medicine **8**(4), 420–422 (2020)
24. Yang, X., He, X., Zhao, J., Zhang, Y., Zhang, S., Xie, P.: Covid-ct-dataset: a ct scan dataset about covid-19. arXiv preprint arXiv:2003.13865 **1**(1), 14 (2020)