International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 10 Issue: 01 | Jan 2023 www.irjet.net p-ISSN: 2395-0072

A Review on Covid Detection using Cross Dataset Analysis

Komal Bangarwa[1], Najme Zehra Naqvi[2]

[1]*PG Student, Computer Science and Engineering, Indira Gandhi Delhi Technical University for Women, New Delhi, India*
[2]*Professor, Dept. of Computer Science and Engineering, Indira Gandhi Delhi Technical University for Women, New Delhi, India*

Abstract - *In the history of the medical field, Covid-19 is one of the deadliest viruses which affected the whole world. The Covid-19 pandemic caused disastrous loss of the human race and unpredictive challenges for community health and even affected the professional world. Various covid detection approaches have been used for fast detection. One of them is medical imaging, specifically Computed Tomography plays a decisive role in examining and monitoring of infected cases. Machine Learning (ML) and Deep Learning (DL) methods also have a significant part in developing different models that can help in diagnosing. This paper will give an overview of deep neural learning approaches used for corona-virus detection by utilizing CT scans in cross-dataset analysis. Even though with ongoing research, model performance has improved with time but still there are areas to cover in cross-dataset analysis. Some limitations observed were the generalization problem, dataset bias problem and robustness which is the capability to work with difficult images that occur due to the variation in image technology. The paper also gives an overview of the methods used so far for pre-processing of medical images and transfer learning techniques. Cross-dataset analysis works on improving the model accuracy by handling the above-mentioned knowledge gaps and providing a model that works more practically by handling different datasets from different sources of the same task and tests how well it adapts to the model. The result might help in understanding the approaches used to overcome the limitations and identify further opportunities in cross-dataset analysis.*

***Key Words*: Adaption, Generalization, covid-19 dataset, Computed Tomography**

1. INTRODUCTION

World Health Organization named corona-virus as Covid-19 which is considered as one of the deadliest contagious diseases that originated from the SARS-CoV virus [14]. The last quarter of year 2019 was the beginning phase of this novel virus which has affected the whole world in every sector either its mass-production, shipment or hospitality. Coronaviruses mostly attack respiratory organs [15]. Most people who were infected by this virus encounter mild to moderate level respiratory discomfort but many recover without requiring any especial medical procedure. However, there exist many cases where the condition was severe and needed more medical surveillance. Geriatric people and the one who already have some inherent medical conditions like diabetes, heart disease, chronic respiratory disease or cancer were at high risk of reaching a serious condition. Many fast covid detection techniques have been used to check the presence of viruses inside the human body like the RT-PCR test[12][13]. But when cases go to a serious level and the oxygen level started to drop below a certain level (~92%), doctors started to advise CT-scan results for the patients infected by Covid. A computerized tomography scan (i.e., CT scan) provides highly precise images of internal body parts, muscles and blood arteries. This allows doctors to recognize the internal function and analyse their shape, thickness and texture. CT scans are better than X-rays because CT scan provides a set of a portion of the particular region without overlapping different body structures. That's why CT scans are better because it gives more detailed information through images which helps in extracting the exact problem and its allocation. A numerous DL methods have evolved and are used for examining Covid-19 by using tomography scans.[24][25][26]. With time, the amount of CT-scan images begins to increase in number and that results in having a good amount of data that can be used for training models that help in better results. Models trained with one dataset should give good accuracy to similar types of datasets which is one of the main factors in determining how well our model learned. Two different datasets are used for cross-dataset analysis in which first is alluded to source dataset which is utilized for training the model & the second is the target dataset which is unseen to the model. Their sources are different but they should address the same task. Many different attributes have been seen in cross-dataset over the years. A short annotation related to its attributes [19] is explained below.

1) Attributes on data:

- *Data accessibility:* This refers to the accessibility and ampleness of both source data and target data.
- *Balanced between data:* Do the provided datasets have a balanced number of data samples?
- *Continuous data:* Is the provided data continuous or online and is it progressing with time?
- *Feature space:* Consistency of identified features in data.

2) Attributes on the label:

- *Available labels:* This refers to the accessibility of labels in both datasets.
- *Label space:* Is classification in both data samples similar?

Considering the above features, different frameworks are being observed.

a) *Feature spaces and label spaces inhomogeneous datasets:* Both datasets should have comparable feature space and label space. But their data generation sources should be different.

b) *Feature spaces in heterogeneous datasets:* The feature space of both datasets should be contrasting which means their source should be diverse but the classification should be the same.

c) *Label spaces in heterogeneous datasets:* The label space of both datasets is contrasting which means their task is different but their features should be alike.

d) *Feature and label spaces in heterogeneous datasets:* The feature and label spaces of both datasets are contrasting which means the source and task are different.

Sometimes because of the complications, many deep neural models require a large number of inputs. However, for the majority, there is a lack of data available for work and for cross-dataset, a good amount of dataset is required for training and testing the model. To overcome, this we can use two methods, first is data augmentation and second is transfer learning.

A. Data Augmentation

The technique helps in increasing the count of training samples by remodeling the images without losing their original knowledge. Remodeling like rotation, horizontal flip, scaling. This method retains images and will help the physician in explaining the images.

B. Transfer learning

Training a new model with the help of a pre-trained network & re-training it with another dataset is referred to as transfer learning [21]. A fine-tuned performance by a pre-trained model can be used for deep neural if there is a scarcity of training data since the model already have knowledge from different problems which can be reused [22] in the current model. There are a few steps that are needed for using transfer learning:

1. Weights used in the pre-trained model are copied to the new build model.
2. Modifications were made in the newer model's architecture, together with the recent layers from the top of the model.
3. Randomly initializing the newer added layers.
4. Defining layers to use for the learning process.
5. Performing the learning phase, by improving the weights.

There is a number of the advanced method being developed using transfer learning over different datasets[20]. Transfer learning not only solves the data scarcity problem but also helps in saving time and improve the performance of the model. In this paper, we will be using cross-dataset analysis on CT scan images used for covid detection. The remainder of the paper is organized as follows: under the literature survey as we discussed the analyzing methods used so far, followed by a conclusion, future scope and ending with references.

2. REVIEW OF LITERATURE SURVEY

In past, many research works have been intensively conducted in machine learning and deep learning fields and through them, many new and improvised methods have been developed in every field. Here we will be understanding the work that has been done in diagnosing Covid-19, a global pandemic. We'll be concisely reviewing the methods used for image classification focusing on CT (Computed Tomography) scans through deep learning techniques.

Below mentioned Table 1, gives a piece of information about the datasets which includes the total images in the dataset with a count on the number of covid and non-covid images in datasets. Table 2, presents which dataset is used against which dataset in studies, named Dataset 1 and Dataset 2.

Table -1: Dataset details

Dataset Information			
Name of the dataset	Total Images	Covid19 Images	Non-Covid19 Images
SARS-CoV-2[1]	2482 CT images	1252 images	1230 images
COVID-CT[2]	746 images	349 images	397 images

Table -2: Cross-Dataset Table

Cross-Datasets Table		
Dataset 1	Dataset 2	Dataset used for training
SARS-CoV-2	COVID-CT	SARS-CoV-2
COVID-CT	Hust-19	Hust-19
SARS-CoV-2 + COVID-CT	MosMed	SARS-CoV-2 + COVID-CT

Many methods have been used in deep learning for image classification and extracting the information that helps in producing the best outcome from a model. Before using the actual method, the provided data has to be prepared according to the need of the model, this process is called pre-processing. It defines the modification process that will be performed on the data before using it in the algorithm. This is important as we can't give raw data, we need to modify the raw and primary data into clean and properly defined data for achieving better results. Many techniques have been used for pre-processing which include, (i) a pixel intensity normalization between the scope from [0, 1][5][7] (ii). Using prepNet[6] helps in standardising CT images concerning the visual differences in datasets before training. One advantage of using prepNet[23] is, this approach can be combined with any diagnosis model, which is a plus point for future progress without additional efforts while it can improve cross-dataset performance. Data augmentation is used for increasing the count of data without losing its knowledge.

Zhao Wang[8][30], uses COVID-Net for rapid identification of Covid-19 but he observed some major restrictions occurred due to the recklessness caused by data heterogeneity in different medical centres regarding numerous imaging environments (for example, imaging protocols, scanner vendors etc.). The two datasets which he has used for his work appears with different image contrast which eventually affects the model outcome. Zhao Wang, redesign the model because the original COVID-Net "lacks in internal feature normalization" [9] which leads to notable variance as CT scan images contain more elaborated patterns, such variance can lead towards the halting of the training stage and influence the prediction accuracy. For handling such variance, batch normalization layers were incorporated [16] to improvise the feature differentiation capability and speed up the training stage. He experimented in two different setups,

1) a single setting, where the model gets trained with every single dataset; and 2) using two datasets (one assigned for training & other for testing). By using redesigned model improvement in results was observed by 12.16% and 14.23% in accuracy. The results which were recorded using the redesign Covid-Net were 96.24% and 85.32% in accuracy (when training the model using the SARS-CoV-CT dataset and Covid-CT dataset, respectively.)

Efficient-Net is a part of a neural network where the "Mobile Inverted Bottleneck Conv Block" [17] is considered the primary block of the network. The main idea of Efficient-Net architecture is to initialise the model using high quality yet compact base model and progressively compare its dimensions, in a structured pattern using an unwavering scaling coefficient. Pedro Silva [5] performed an experiment which is one of the earliest performed cross-dataset analyses. He uses Efficient-Net, where he also performed data augmentation and uses transfer learning. His team uses the SARS-CoV-CT dataset for training the model and has kept other datasets Covid-CT, completely for testing the model. From his work, he observed that there is a drop in accuracy as compared to individual dataset accuracy and it got worse when the experiment was reversed i.e., training with Covid-CT and testing on SARS-CoV-CT, showing the accuracy drop from 87.68% to 56.16%. They observed that the reason was the lack of diversity in images and image technology issues.

Haikel Alhichri[10] uses Efficient Net-B3 which is a bit faster than Efficient-Net. He used his own developed pre-processing method to resize all image data to 256 x 256 pixels which will reduce image loss. Formerly "the EffecientNet-B3 model is pre-trained on the ImageNet dataset [18]". He modified Efficient-Net-B3 to Efficient-Net-B3-GAP (Global Average pooling). He uses the modified version because it has more relevance to the convolution structure by building a correlation between feature maps and categories. The resulting outcome when trained with SARS-CoV-CT and COVID-19 datasets respectively are 99.1% and 88.18%. His observations state that adaptation is needed to get better results.

Fozia Mehboob [11]uses ViT, a self-attention mechanism. It was observed that the Vision transformer yielded better results than convolutional operations [27]. Although the major advantage of CNN is that it provides a comparison to existing classification but CNN architecture is specific to the task and its computational time is more. A convolutional network uses the information processed by pixel where every pixel has its own significance for the target task which causes reiteration and consumes more time. ViT, is a vision model which takes input images as image patches. The following patches are further fattened for assigning it to the transformer encoder. In his study, he uses a new dataset called Hust19 with SARS-CoV-2 dataset for analysing. He trained the model with the Hust19 dataset and tested it with the SARS-CoV-2 dataset. He also evaluated another scenario where the SARS-CoV-2 dataset has been used for training and the Hust19 dataset for testing. In the second scenario, a downfall in model performance is observed and according to his observation, the possible reason could be the variation in images. Dissimilarity among images was higher in comparison with other datasets and the number of image sample associated with each variation were quite low. The accuracy achieved using Hust19 as the training dataset was 93%. Since ViT (vision transformer) performed poorly in generalization because of the small dataset this could be one

of the reasons for the decline in accuracy. When the same model is used with the Hust19 and Covidx datasets the accuracy was comparatively good because the images in both datasets are similar. The achieved accuracy was 94% which was higher when compared to individual datasets.[28][29].

Table -3: Summary of Literature Survey

Review table			
Literature	Dataset	Method	Result
Mohammadreza Amirian et al. 2022	Public datasets (SARSCOVID-2, UCSD COVID-CT, MosMed	Autoencoders(Vgg Net-16), PrepNet. Trained on SARSCOVID-2 and COVID-CT. Tested on MosMed.	PrepNet improves the cross-dataset balanced accuracy by a margin of 11.84 percentage points
Zhao Wang et al. 2020	Covid-CT dataset and SARS-CoV-2 CT Scan dataset	COVID-Net Trained On SARS-CoV-2 CT data and Tested on Covid-CT data.	The accuracy achieved 85.6% whereas the performance score was increased by 7.8%.
Pedro Silva et al. 2020	Covid-CT dataset and SARS-CoV-2 CT Scan dataset	Efficient-Net model. Trained on SARA-CoV-2-CT and tested on Covid-CT	This was the first time this experiment was done so accuracy from 87.86% dropped to 59%.
Haikel Alhichri et al. 2020	Covid-CT dataset and SARS-CoV-2 CT Scan dataset	Two models used: DenseNet & EfficientNet B-3(with Vgg-16, ResNet) Trained on SARS-CoV-2 CT Scan and Tested on Covid-CT	Achieved an accuracy of 96.4%.
Ludmila Silva et al. 2020	Covid-CT dataset and SARS-CoV-2 CT Scan dataset	EfficientNet model. Trained on Covid-CT and tested on SARA-CoV-2-CT. Voting-based approach.	Achieved an accuracy of 88%.
Fozia Mehboob et al. 2022	SARS-CoV-2 CT Scan dataset, HUST-19, CovidX.	Vision Transformer (ViT). Training on the Hust19 dataset and testing COVIDx dataset.	Achieved an accuracy of 94%.

3. CONCLUSIONS

Due to the shortage and false alarms by RT-PCR test and the drastic increase in Covid-19 cases the tomographic scan together with DL approaches proved to be very useful in providing fast and highly accurate results. The model performs better when it gets trained by a large quantity of data using DL techniques. Hereby we came across the cross-dataset analysis concept and some methods which helps in evaluating some knowledge gaps observed during the evaluation. Cross-dataset analysis is not a common method and there is a lot to unwrap in this process. The most common limitations that were observed were the generalization process which deals with the capability of the model to work with independence and robustness which deals with the capacity of a model to work with difficult images and the dataset bias problem, which tackles the problem that occurred because of identical data which causes loss when a new data is given to the model. We have discussed pre-processing method used to enhance the properties of medical images. Due to insufficiency in the availability of data volume used for training, maximum models have used data augmentation and transfer learning approaches. Most of the DL methods for Covid-19 detection use pre-trained models to streamline the process during the quick increase in cases. Efficient-Net, Covid-Net, Vision transformation(ViT), and DenseNet are the pre-trained models mostly used in detection. The role of the medical imaging technique, CT scan is also briefly discussed. CT images are highly sensitive for analysing the seriousness of the covid condition.

As it is a new pandemic, more research can be done using a large number of datasets obtained from different sources. Results came up from various trained transfer learning models that could be either incorporate to build up a hybrid model or could be used for comparison when systematic changes are made in models to improve accuracy. In future, research can focus on getting more improvised methods for improving the quality of images and the model that works on different datasets, providing only one dataset for training to deal with generalization problems and problems occurred due to imaging technologies. This will help in diagnosing the covid-19 more accurately even by providing CT scan images from a completely new environment.

REFERENCES

[1] E. Soares, P. Angelov, S. Biaso, M. Higa Froes and D. Kanda Abe, "Sars-cov-2 CT-scan dataset: A large dataset of real patients CT scans for sars-cov-2 identification", medRxiv, 2020.

[2] J. Zhao, X. He, X. Yang, Y. Zhang, S. Zhang and P. Xie, "Covid-CT-dataset: A CT scan dataset about covid-19", 2020.

[3] Hust-19dataset: "https://bioengineeringcommunity.nature.com/posts/hust-19-for-predicting-covid-19-clinical-outcomes"

[4] 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.

[5] Pedro Silva, Eduardo Jose da Silva Luz, Guilherme Silva, Gladston Moreira, "Efficient Deep Learning Model for COVID-19 Detection in large CT images datasets: A cross-dataset analysis" 2020

[6] Mohammadreza Amirian, Javier A. Montoya-Zegarra, Jonathan Gruss, Yves D. Stebler, Ahmet Selman Bozkir, Marco Calandri, Friedhelm Schwenker, Thilo Stadelmann, "PrepNet: A Convolutional Auto-Encoder to Homogenize CT Scans for Cross-Dataset Medical Image Analysis",2022.

[7] Pedro Silva, Eduardo Jose da Silva Luz, Guilherme Silva, Gladston Moreira," COVID-19 detection in CT images with deep learning: A voting-based scheme and cross-datasets analysis", 2020.

[8] Zhao Wang, Quande Liu, Qi Dou," Contrastive Cross-Site Learning With Redesigned Net for COVID-19 CT Classification, 2020

[9] L. Wang and A. Wong, "Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images", 2020.

[10] Haikel Alhichri, "CNN Ensemble Approach to Detect COVID-19 from Computed Tomography Chest Images", 2021

[11] Fozia Mehboob, Abdul Rauf, Richard Jiang, Abdul Khader Jilani Saudagar, Khalid Mahmood Malik, Muhammad Badruddin Khan, Mozaherul HoqueAbdul Hasnat, Abdullah Al Tameem, Mohammed AlKhathami. ''Towards robust diagnosis of COVID‑19 using vision self‑attention transformer''

[12] Organization, W.H (2020) Clinical management of severe acute respiratory infection when novel coronavirus (2019-nCoV) infection is suspected: interim guidance, in the clinical management of severe acute respiratory infection when novel coronavirus (2019-nCoV) infection is suspected: interim guidance

[13] Alsharif W, Qurashi A. "Effectiveness of COVID-19 diagnosis and management tools: a review." Radiography. 2020;27:682–687. DOI: 10.1016/j.radi.2020.09.010.

[14] Kristian G. Andersen, et al. "The proximal origin of SARS-CoV-2 Nature medicine",2020.

[15] Tanu Singhal, "A review of coronavirus disease-2019(COVID-19)", 2020

[16] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift", arXiv:1502.03167, 2015.

[17] Zhao, Y. Zhang, X. He, P. Xie, "Covid-ct-dataset: a ct scan dataset about covid-19", arXiv preprint arXiv:2003.13865 (2020).

[18] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh et al., "ImageNet large scale visual recognition challenge," International Journal of Computer Vision, vol. 115, no. 3, pp. 211–252, 2015.

[19] Jing Zhang, Wanqing Li, Philip Ogunbona,Dong Xu "Recent Advances in Transfer Learning for Cross-Dataset Visual Recognition: A Problem-Oriented Perspective", 2019

[20] Sinno Jialin Pan and Qiang Yang, A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering 22, 10 (2010), 1345–1359.

[21] Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, Karpathy, A. Khosla, M.Bernstein, et al., "Imagenet large scale visual recognition challenge", International journal of computer vision 115 (3) (2015) 211–252. Goodfellow, Y. Bengio, A. Courville, Deep learning, MIT Press, 2016.

[22] Oquab, L. Bottou, I. Laptev, J. Sivic, Learning and transferring mid-level image representations using convolutional neural networks, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 1717–1724.

[23] 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 covid19 detection using ct scans with detail-oriented capsule networks," arXiv preprint arXiv:2004.07407, 2020.

[24] He X., Yang X., Zhang S., Zhao J., Zhang Y., Xing E., Xie P. medRxiv; 2020. Sample-efficient deep learning for covid-19 diagnosis based on ct scans.

[25] Polsinelli M., Cinque L., Placidi G. 2020. A light cnn for detecting covid-19 from ct scans of the chest

[26] Amyar A., Modzelewski R., Ruan S. medRxiv; 2020. Multi-task deep learning based ct imaging analysis for covid-19: classification and segmentation

[27] Transformers for Image Recognition at Scale—Google AI Blog. Dec 2020, http://ai.googleblog.com

[28] Ning, W. et al. Open resource of clinical data from patients with pneumonia for the prediction of COVID-19 outcomes via deep learning. Nat. Biomed. Eng. 4(12), 1197–1207 (2020).

[29] Zheng, C., Deng, X., Fu, Q., Zhou, Q., Feng, J., Ma, H., Liu, W., Wang, X. (2020) Deep learning-based detection for COVID-19 from chest CT using a weak label. MedRxiv.

[30] Wang L, Lin ZQ, Wong A. "Covid-net: a tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images". Sci Rep. 2020