Pakistan Journal of Engineering and Technology, PakJET
ISSN (p): 2664-2042, ISSN (e): 2664-2050
Volume: 5, Number: 2, Pages: 73- 78, Year: 2022

Comparative Analysis of Different Machine Learning Models Applied on Different Datasets Available for Diagnosis of Covid-19

Osama Ali[1], Adeel Muzaffar Syed[1], Joddat Fatima[1] and Samina Khalid[2]

[1]Department of Software Engineering, Bahira University, Islamabad, 44000, Pakistan
[2]CS and IT Department, Mirpur University of Science and Technology, Pakistan

Corresponding author: Adeel Muzaffar Syed (e-mail: adeel@bahria.edu.pk).

Received: 15/01/2022, Revised: 20/05/2022, Accepted: 15/06/2022

Abstract- **Since breaking out in Wuhan, China in the last days of 2019, the novel COVID-19 pandemic has done a great deal of damage to mankind; whether it is economic damage, psychological or social. It was declared a pandemic in March 2020. PCR (Polymerase Chain Reaction) which is used to diagnose COVID-19 patients usually takes 24-72 hours ranging in different countries. A new idea of diagnosing COVID-19 with the help of radiography images has surfaced which has taken research world by storm. There are different machine learning models developed with the help of historic data i.e., datasets that classify COVID-19 patients within a few minutes. As there are different publicly available datasets on which dozens of models are developed, we would like to perform a comparative analysis of these datasets. This would help us to identify different aspects of these datasets.**

Index Terms-- **COVID-19, CNN, CT scans, CXR, PCR, Pneumonia, Radiography.**

I. INTRODUCTION

A newly known disease, affecting living species of the earth, known as COVID-19 emerged in December 2019 [1]. Its impact was so devastating that soon it was declared a pandemic in March 2020 by WHO. Weekly, millions of cases started surfacing according to COVID world-o-meter data [2]. Since then, intellectual people and different administrations all over the world are trying to curb this. Lockdowns were part of this exercise, but this was a temporary solution. Medical Fraternity developed a COVID-19 test mechanism called PCR (Polymerase Chain Reaction) [3] for diagnosing COVID-19 patients but another problem was around the corner. PCR test at best used to take 24 hours to produce results. During this time, the patient would have contacted more people and this chain was going alone. The amount of time PCR test takes had led scientists to think differently.

To address this problem, a new idea of diagnosing through radiography images [4] was presented. Machine learning algorithms are applied to radiography image datasets of COVID-19 patients started being used to classify COVID-19 patients. The good thing about these frameworks was that it normally used to take 30-60 minutes for diagnosis. Moreover, most models produced an accuracy of an excess of 90% but still, there were a few problems.

In this research, we will talk about these problems regarding these models as well as datasets. A comparative analysis is required to find out all the points. This can help us to identify the hidden aspects of these models as well as datasets. These aspects include positive as well as negative points. Normally, researchers don't have time to carefully analyze these aspects. Moreover, comparative analysis helps us to broaden our thinking about product functionality as well as relevant patterns are identified [5].

A. ABBREVIATIONS AND ACRONYMS

C.N.N. stands for Convolutional Neural Network. C.T. means Computed Tomography. C.X.R. means Chest X-Ray. OptCoNet stands for Optimized Convolutional Neural Network

II. LITERATURE REVIEW

Hussain et al. [6] proposed a deep learning model which classifies COVID-19 patients using chest X-ray images. He used a 22-layer CNN (Convolutional Neural Network) Architecture named GoogleNet for classification. This model is applied to an X-ray dataset. This model is claimed to have performed different classifications i.e., 2 class classifications (COVID and Normal), 3 class classifications (COVID, Normal and Pneumonia) and 4 class classifications (COVID, Normal, Viral Pneumonia and Bacterial Pneumonia). This produced different results for

different classification problems. For 2-class problem, its accuracy was recorded as 99.1%, 94.2% for 3-class and 91.2% for 4-class problem. The results showed this Model had better results than the other 10 existing MODELS.

Asif et al. [7] presented a model for the classification of COVID-19. They used digital x-rays for this purpose. They applied CNN architecture Inception V3. However, they performed only 3 class classification. They used a dataset having images of around 3500 which included 864 COVID-19 images, 1345 viral pneumonia and 1341 normal x-ray images. This Model produced an accuracy of 98%.

Mahdy et al. [8] proposed technique for COVID-19 classification with the help of lung images. The proposed used multi-level thresholding procedure and SVM (Support Vector Machine) Algorithm. All the images used in this dataset were of the same size, same format (JPEG) and same pixel (512*512). This sensitivity, specificity, and accuracy for this model were recorded as 95.76%, 99.7%, 97.48% respectively. One of the drawbacks of this Model was, that it was 2 class classification.

Lujan-Garcia et al. [9] proposed a model which aims to perform COVID-19 diagnosis at a faster rate and Pneumonia classification images by using 2 different datasets. One dataset consisted of pneumonia-inflected CXR images and the other consisted of Normal and COVID-19 images. This diversity of this model was unique as the dataset had records ranging from 12 to 87 years, so it covered all the range of humans. So, CNN algorithms were applied to these datasets. Performance evaluation metrics confusion metrics were used to determine the performance of the model. The results showed that the performance was excellent as results were obtained in 12 minutes.

Narin et al. [10] presented a model with the same attributes but with a different approach. 5 different CNN Architectures were used in this study namely ResNet50, ResNet101, ResNet152, Inception V3 and Inception-ResNetV2. It performed 4-class problems (COVID, Normal, Viral Pneumonia and Bacterial Pneumonia) by using five-fold cross-validation. The good thing about this work was that it was applied to three different datasets. The researchers used 70% of the data for the training, 10% for validation, and the remaining 20% for testing.

Kumar et al. [11] proposed a deep learning model with machine learning classifiers and SMOTE. SMOTE is a feature used in python to imbalanced data [12]. It uses K-nearest neighbor algorithm to bring synthetics into data [13]. This model achieved an accuracy of 97.3% on Random Forest and 97.73% on XG Boost predictive classifiers [14].

TABLE I
AN OVERVIEW OF THE MODELS USED BY DIFFERENT RESEARCHERS

S. No	Author	Year	Diseases	Dataset	Results
1	Hussain et al. [6]	2021	COVID-19, Normal Pneumonia, Bacterial Pneumonia	Kaggle [25]	More than 90% Accuracy for all classification
2	Asif et al. [7]	2020	COVID-19, Pneumonia	Github [21]	98% Accuracy
3	Mahdy et al. [8]	2020	COVID-19	Kaggle [22]	95% Sen, 99% Spe, 97% Acc
4	Luján et al. [9]	2020	COVID-19	Github [21] Mendeley [28]	More than 90% Accuracy
5	Narin et al. [10]	2021	COVID-19, Normal Pneumonia, Bacterial Pneumonia	Paper-with-code [27]	More than 90% Accuracy
6	Kumar et al. [11]	2020	COVID-19	Kaggle [14] towards data [29]	More than 90% Accuracy
7	Thakur et al. [15]	2021	COVID-19, Pneumonia	Kaggle [22]	More than 90% Accuracy
8	Anter et al. [16]	2021	COVID-19	Github [21]	More than 90% Accuracy
9	Narin et al. [17]	2021	COVID-19, Pneumonia	Kaggle [22]	99% Accuracy
10	Goel et al. [18]	2021	COVID-19	Github [21]	97% Accuracy
11	Karthik et al. [19]	2021	COVID-19	Kaggle [14]	99% Accuracy

Dhiman et al. [20] suggested the DON framework for the detection of novel COVID-19 disease using X-ray images. Eleven different frameworks were used in this proposed framework. These eleven Models were AlexNet, VGG16, VGG19,

GoogleNet, ResNet, ResNet18, ResNet500, ResNet101, Inception V3, InceptionResNetV2, DenseNet201, XeptionNet. The model was claimed to have outperformed competitive models in the race of precision, recall, accuracy, specificity, and F1 Score [14].

Thakur et al. [15] proposed a methodology in which the classification of COVID-19 was performed by applying the CNN Algorithm. They divided their work into 2 parts. First, they developed a model for two-class classification, and then they developed another Model for multiple class classification. For binary classification, a dataset containing more than 3800 images was used of which 1917 CXR images belonged to COVID infected patients. Similarly, for multiple class problems, more than 6000 images were used out of which 1917 were COVID infected images. All the commonly used parameters of a model were giving results of more than 90%.

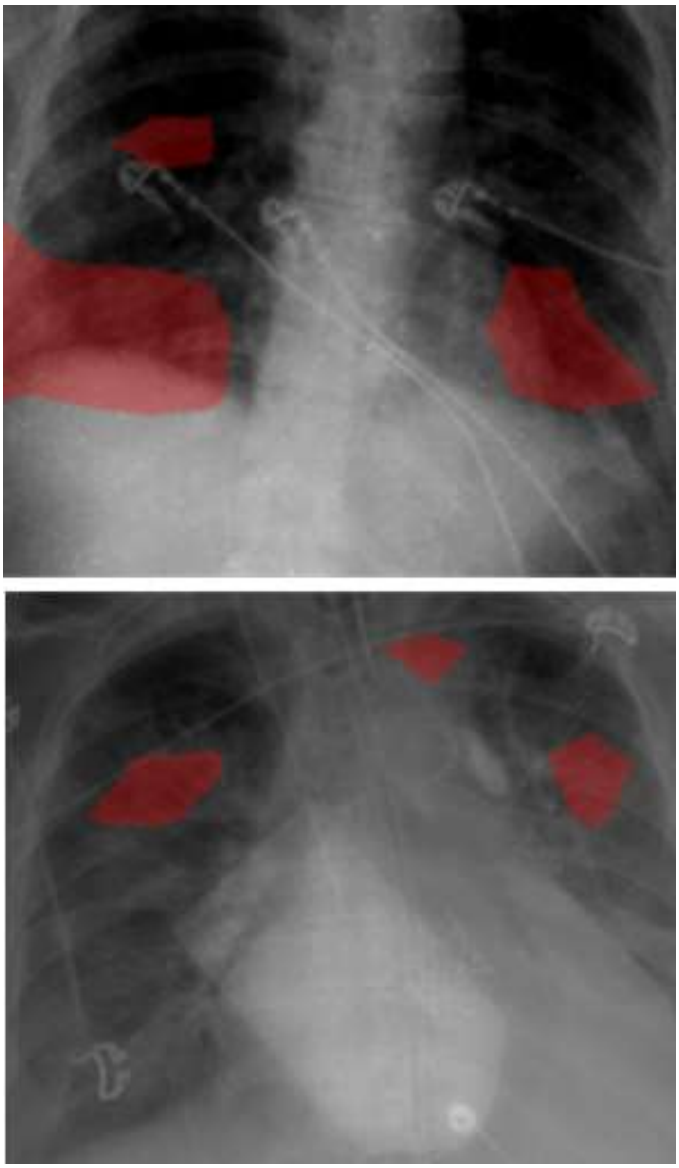

FIGURE 1. Sample Images of 2 Different Patients from COVID-NetCXR-2 Dataset and Their Associated Critical Factors

Anter et al. [16] presented a model named AFCM-LSMA for the detection of COVID-19. This is a new intelligent model based on the combination the of algorithms Fuzzy C-Means and Slime Moulid Algorithm. This model produced an accuracy, Precision, and F1 Score of 96%, 98%, and 98% respectively [21].

Narin [17] used Deep Learning and old machine learning methods were used together to develop a model for detection of COVID-19 patients. It was implemented with three different CNN architectures (ReSNet50, ResNet101, and InceptionResNetV2). It was used for a three-class problem i.e., COVID-19, Normal and Pneumonia. The accuracy for this Model was recorded at 99.86%.

Goel et al. [18] presented a paper that aims to detect COVID-19 patients through CNN Algorithm. A Framework OptCoNet is proposed. GWO Algorithm was also used in this framework. This model is trained on 1890 images and then tested 810 images. It provided an accuracy of 97.78%.

Karthik et al. [19] suggested a custom CNN architecture which learns unique convolutional filter patterns. This model also determines saliency of X-ray images. This model produced an accuracy of 99.8% [14].

We have summarized these results in table 1 below where authors, year published, a general review of the technique used, diseases, datasets and evaluation metric are discussed

III. DATASETS

Dataset is a collection of data. It may consist of numerical data, alphabetic data or even image data. We may refer it as small database where data is stored in files and folders. There is always a common element i.e., pattern in data of one dataset.

In this section, we are taking different datasets of radiography images. This dataset would be used in implementing different models through machine learning algorithms which would be used to perform COVID-19 classification.

A. COVIDx CXR-2 DATASET

COVIDx CXR-2 is a publicly available dataset [22] containing over 16000 Chest X-ray images. It is ideal dataset for training and testing Model our COVID-19 detection. This was created by Andy Zhao around the start of COVID-19 pandemic. At the start, it consisted of few hundred images, but it kept increasing as time went on. A sample of the COVIDx CXR-2 is shown in Fig 1.

B. COVIDx CT

This dataset is created by Hayden Gunraj [23]. This is an openly accessible dataset consisting of CT images. This dataset is a combination of different publicly available datasets. It is also used train and test COVID-19 detection models. The difference is that this model is developed with the help of CT images. Fig 2 shows a sample image of this dataset.

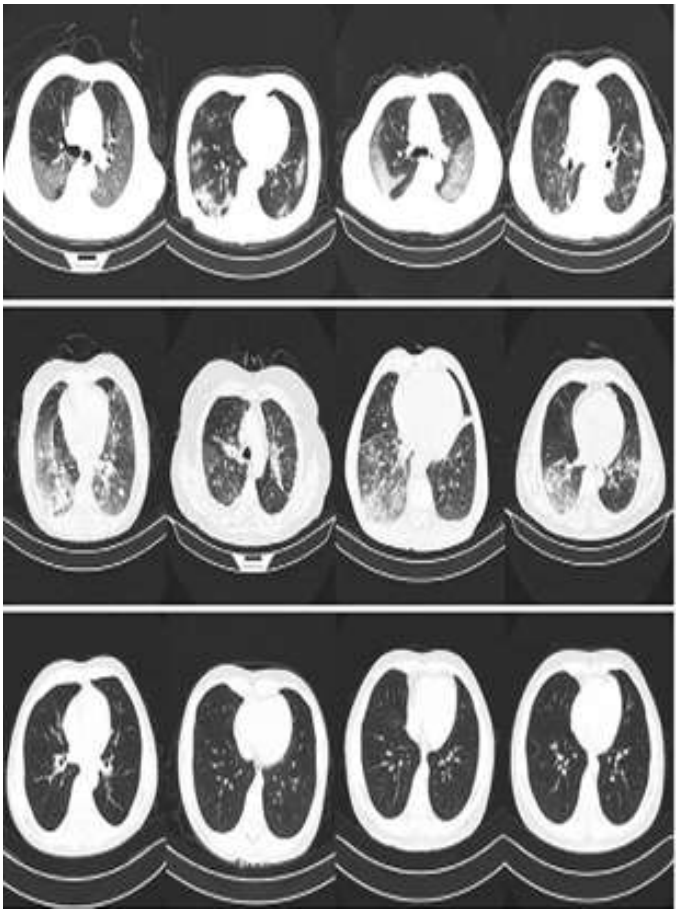

FIGURE 2. Chest CT images from the COVIDx-CT dataset, illustrating (A) COVID-19 pneumonia cases, (B) non-COVID-19 pneumonia cases, and (C) normal control cases

C. COVID-19 RADIOGRAPHY DATABASE

COVID-19 Radiography Database was developed by a team of researchers from different countries mainly by Qatar University, Doha and University of Dhaka, Bangladesh [24]. They are continuously updating this database which have now over 20000 images. It consists of lung opacity, normal and pneumonia images. Fig 3 shows a sample of this dataset.

D. PNEUMONIA X-RAY IMAGES

Pneumonia X-Ray Images was developed by Paulo Breviglieri [25]. It consists of around 5500 images. It is also used for training and validation purposes in machine learning. A sample image of this database is shown in Fig 4. A summary of these datasets is presented in Table II where the name of the dataset is followed by the sponsor of the dataset. We have also provided the Type of images in the dataset along with the number of images present.

FIGURE 3. A random image of COVID-19 Radiography Database dataset

FIGURE 4. An image of Pneumonia X-Ray Images dataset showing human X-ray

TABLE II
DATASET DETAILS

S. No	Name	Sponsor	Type	Format	Area	Images
1.	COVIDx CXR-2	University of Waterloo, Canada [22]	Chest X-ray	480*480	COVID-19	16,000+
2.	COVIDx CT	Frontiers in Medicine [23]	CT scans	512*512	COVID-19	194,000 +
3.	COVID-19 Radiography Database	Qatar University, Doha-2713, Qatar [24]	Chest X-ray, CT scans	299*299	COVID-19, Pneumonia	20,000+
4.	Pneumonia X-Ray Images	Attribution 4.0 International [25]	Chest X-ray	1024*1024	COVID-19, Pneumonia	5,500+

IV. CONCLUSION

As we explained above, earlier diagnosis is helpful in a pandemic like COVID-19 to help us avoid spreading the disease [7]. Also, it can help to reduce the burden on Health system. In this study, we analyzed different machine learning models developed for diagnosis of COVID-19. We discussed their mechanism and their way of function. We also discussed few datasets and their quantity. Another thing that we learnt was the diversity of machine learning models and their effectiveness in the field of Medical and Epidemiology [26].

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

ACKNOWLEDGMENT

We would like to thank our departments in universities: Software Engineering Department, Faculty of Engineering Sciences, Bahria University Islamabad and CS & IT Department, Mirpur University of Science and Technology for providing ample support for our research. We are thankful to the PG program of Bahria University which supports the research culture for MS students and motive them for publication in their degree.

REFERENCES

[1] Guardian, “First Covid-19 case happened in November, China government records show – report” Jan 12. 2022. [Online]. Available: https://www.theguardian.com/world/2020/mar/13/first-covid-19-case-happened-in-november-china-government-records-show-report

[2] WHO, “World o meter, COVID Live Data." Jan 12, 2022. [Online]. Available: https://www.worldometers.info/coronavirus

[3] L. Garibyan and N. Avashia, “Research techniques made simple: polymerase chain reaction (PCR),” *The Journal of investigative dermatology*, vol. 133, no. 3, pp.e6, 2013.

[4] Chen, S.G., Chen, J.Y., Yang, Y.P., Chien, C.S., Wang, M.L. and Lin, L.T., “Use of radiographic features in COVID-19 diagnosis: Challenges and perspectives,” *Journal of the Chinese Medical Association*, vol. 83, no. 7, pp 644, 2020.

[5] Dorn, B. "Advantages of Comparative Analysis." Apr 27, 2021. [Online]. Available https://www.viget.com/articles/the-advantage-of-comparative-research/. [Accessed: 01-01-2022]

[6] E. Hussain, M. Hasan, M. A. Rahman, I. Lee, T. Tamanna and M. Parvez, “CoroDet: A deep learning based classification for COVID-19 detection using chest X-ray images,” *Chaos, Solitons & Fractals*, vol. 142, pp.110495, 2021.

[7] S. Asif, Y. Wenhui, H. Jin and S. Jinhai (2020).” Classification of COVID-19 from Chest X-ray images using Deep Convolutional Neural Network,” in Proc. *IEEE 6th International Conference on Computer and Communications (ICCC)*, 2020, pp 235-433.

[8] L. N. Mahdy, K. A. Ezzat, H. H. Elmousalami, H. A. Ella and A. E. Hassanien. "Automatic X-ray COVID-19 Lung Image Classification System based on Multi-Level Thresholding and Support Vector Machine," *MedRxiv*, 2020.

[9] J.E. Luján-García, M.A. Moreno-Ibarra, Y. Villuendas-Rey, and C. Yáñez-Márquez. “Fast COVID-19 and pneumonia classification using chest X-ray images,” *Mathematics*, vol. 8 no. 9, pp 1423, 2020.

[10] Narin, A., Kaya, C. and Pamuk, Z., 2021. “Automatic detection of coronavirus disease (covid-19) using x-ray images and deep convolutional neural networks,” *Pattern Analysis and Applications*, vol. 24, no. 3, pp.1207-1220, 2021.

[11] R. Kumar, R. Arora, V. Bansal, V. J. Sahayasheela, H. Buckchash, J. Imran, N. Narayanan, G. N. Pandian and B. Raman. "Accurate Prediction of COVID-19 using Chest X-Ray Images through Deep Feature Learning model with SMOTE and Machine Learning Classifiers," *MedRxiv*, 2020.

[12] J. Brownlee. "SMOTE for Imbalanced Classification with Python." Jan 12, 2021. [Online]. Available: https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/.

[13] C. Y. Wijaya. "5 SMOTE Techniques for Oversampling your Imbalance Data." Feb 24, 2021. [Online]. Available: https://towardsdatascience.com/5-smote-techniques-for-oversampling-your-imbalance-data-b8155bdbe2b5

[14] P. Mooney, "Chest X-Ray Images (Pneumonia)." Jan 14, 2022. [Online]. Available: https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia.

[15] S. Thakur and A. Kumar, "X-ray and CT-scan-based automated detection and classification of covid-19 using convolutional neural networks (CNN)," *Biomedical Signal Processing and Control*, vol 69, pp. 102920, 2021.

[16] A. M. Anter, D. Oliva, A. Thakare and Z. Zhang, "AFCM-LSMA: New intelligent model based on Lévy slime mould algorithm and adaptive fuzzy C-means for identification of COVID-19 infection from chest X-ray images," *Advanced Engineering Informatics*, vol. 49, pp. 101317, 2021.

[17] A. Narin, "Accurate detection of COVID-19 using deep features based on X-Ray images and feature selection methods," *Computers in Biology and Medicine*, vol. 137, pp. 104771, 2021.

[18] T. Goel, R. Murugan, S. Mirjalili and D. K. Chakrabartty, "OptCoNet: an optimized convolutional neural network for an automatic diagnosis of COVID-19." *Applied Intelligence*, vol. 51 no. 3 pp. 1351-1366, 2021.

[19] R. Karthik, R. Menaka and M. Hariharan, "Learning distinctive filters for COVID-19 detection from chest X-ray using shuffled residual CNN," *Applied Soft Computing* vol. 99, pp.106744, 2021.

[20] G. Dhiman, V. Vinoth Kumar, A. Kaur and A. Sharma. "DON: Deep Learning and Optimization-Based Framework for Detection of Novel Coronavirus Disease Using X-ray Images," *Interdisciplinary Sciences: Computational Life Sciences*, vol. 13, no. 2, pp. 260-272, 2021.

[21] IEEE8023. "GitHub - ieee8023/covid-chestxray-dataset: We are building an open database of COVID-19 cases with chest X-ray or CT images," Oct 10, 2021. [Online]. Available: https://github.com/ieee8023/covid-chestxray-dataset

[22] A. Zhao, "COVIDx CXR-2 Dataset." Oct 10, 2021. [Online]. Available: https://www.kaggle.com/andyczhao/covidx-cxr2

[23] H. Gunraj, A. Sabri, D. Koff and Alexander Wong. "COVID-Net CT-2: Enhanced deep neural networks for detection of COVID-19 from chest CT images through bigger, more diverse learning," *Frontiers in Medicine*, vol 8, 2021.

[24] T. Rehman, "COVID-19 Radiography Database." Oct 10, 2021. [Online]. Available: https://www.kaggle.com/tawsifurrahman/covid19-radiography-database

[25] P. Breviglieri, "Pneumonia X-Ray Images." Oct 10, 2021. [Online]. Available: https://www.kaggle.com/pcbreviglieri/pneumonia-xray-images

[26] T. L. Wiemken and R. R. Kelley. "Machine Learning in Epidemiology and Health Outcomes Research." *Annual Review of Public Health*, vol. 41 pp. 21-36, 2019.

[27] Wang, "ChestX-ray8." Oct 10, 2021. [Online]. Available: https://paperswithcode.com/dataset/chestx-ray8

[28] D. Kermany, M. Goldbaum, "Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification." Oct 10, 2021. [Online]. Available: https://data.mendeley.com/datasets/rscbjbr9sj/2

[29] A. Ketari, "COVID-19 public dataset from cases in Italy on Google Cloud Platform." Oct 10, 2021. [Online]. Available: https://towardsdatascience.com/covid19-public-dataset-on-gcp-nlp-knowledge-graph-193e628fa5cb?gi=a534ead97b81