INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Check for updates

OPEN ACCESS

Received: 19-12-2023

Accepted: 19-06-2024

Published: 13-07-2024

Citation: Almuafi NA, Alzahrani ME (2024) Machine Learning-Based Prediction Approach for COVID-19 Detection via Chest X-Ray Scans. Indian Journal of Science and Technology 17(27): 2813-2819. https://doi.org/10.17485/IJST/v17i27.3126

* **Corresponding author**.

Msnajwa7x@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Machine Learning-Based Prediction Approach for COVID-19 Detection via Chest X-Ray Scans

Najwa Anwar Almuafi[1]*, **Mohammad Eid Alzahrani**[2]

1 M.Sc. Artificial Intelligence, Al-Baha University, Al-Baha, Saudi Arabia
2 Department of Computer Science, Faculty of Computing & Information Al-Baha University, Al-Baha, Saudi Arabia

Abstract

Objectives: In medical research, the application of AI technologies has become a desirable approach for disease diagnosis due to their ability to increase diagnosis speed and accuracy. One major application of these technologies is in diagnosing diseases using medical imagery data such as X-ray and CT scan images. The objective of this study is to implement a COVID-19 diagnosis approach based on ML and compare it with laboratory testing approaches. **Methods:** The study utilizes the VGG16 architecture to extract features from X-ray and CT scan images, which are then used to build an ANN classifier to predict images as either COVID-positive or normal. Real-world datasets of varying difficulty were employed to evaluate the system's performance under different stressors, while comparative analysis with other classifiers was conducted to assess the model's improvement in accuracy and time complexity. **Findings:** The proposed architecture correctly classified 98.41% of the testing images. Among the images from infected patients in the testing set, 97.54% were accurately identified, while 98.72% of the normal cases were correctly classified. The model demonstrated good precision, with 96.49% of predicted COVID cases truly being from COVID-19 patients. Additionally, the F1 score for the model is 98.41%, indicating a good balance in predicting both categories. Showcasing its ability to handle multiple predictions quickly. The model's execution time for classifying a single case was computed as 0.0039 seconds, showcasing its ability to handle multiple predictions quickly. The proposed approach exhibits superior performance on datasets with balanced outcomes, surpassing the accuracy of SVM, DT, Logistic Regression, and Random Forest models. The approach outperforms laboratory-based testing approaches in terms of accuracy cost and assay times. Novelty: This study not only implements a robust ML-based approach for COVID-19 diagnosis but also provides a unique comparative analysis with traditional laboratory-based testing methods, such as Reverse Transcription Polymerase Chain Reaction (RT-PCR), reported in other literature. This dual focus on ML performance and its comparison with established testing methods highlights the practical advantages of ML in terms of diagnosis speed, accuracy, and cost-efficiency, offering a comprehensive perspective not commonly addressed in existing research.

Keywords: COVID-19; Machine Learning techniques; F1 score; Classifier; Specificity; Accuracy; Sensitivity; Precision

1 Introduction

The emergence of the COVID-19 disease disrupted every sphere of life and caused a massive loss of life worldwide. At the onset of the disease, diagnostic techniques relied on epidemiological history, clinical symptom analysis, and positive pathogenic tests[(1)]. However, the rapid transmission of COVID-19 rendered most of these diagnostic approaches ineffective due to cost and waiting time constraints. Consequently, extensive research focused on leveraging advanced technology to diagnose COVID-19 efficiently. The majority of studies suggest that machine learning presents a viable solution capable of delivering diagnosis results promptly, meeting speed and cost requirements, and aiding in the identification of COVID-19 patterns to initiate intervention mechanisms[(2)].

Moreover, as highlighted in[(3)], human practitioners, including doctors and nurses, may get tired and bored with the monotony of doing the same process repeatedly, leading to potential misdiagnoses. Machine learning models can help overcome these shortcomings in conducting COVID-19 screening and diagnostic tests due to the complexity of the required details, particularly in handling the complexity of patient data categorization and interpretation for effective pandemic intervention strategies. A research by[(4)] and COVID-19 diagnosis by applying a machine learning model can assist healthcare practitioners in arriving at sound decisions.

In the diagnosis of chest ailments such as pneumonia, tuberculosis, pneumoconiosis, and SARS, radiographic images serve as a common source of features that were previously manually examined and recently analyzed using machine learning algorithms to diagnose the diseases[(5)]. Owing to the low cost and non-invasive nature of radiographic images, analyzing them with ML models can result in a cheap, fast, and accurate COVID diagnosis approach, which matches the rapid transmission of the disease. Radiographic images are classified into two major types: Computed Tomography (C.T.) and chest radiography (X-ray). These two techniques are used due to their ability to record digital images of the lungs. C.T. images have a lower misdiagnosis rate due to their high sensitivity. C.T. images are preferred for diagnosing chest infections since they provide 3D view images of the lungs. These 3D images are usually converted into 2D sagittal, axial coronal views, allowing for improved diagnosis[(3)].

A review of related literature reveals successful implementations of COVID diagnosis using ML approaches with features extracted from radiographic images, as documented in studies such as[(3,6–13)]. Many researchers have focused on implementing machine learning approaches on radiographic images to diagnose COVID, with the aim of improving diagnosis cost and accuracy, However, there is little or no reference to previously existing traditional diagnosis approaches in these studies. Furthermore, the literature often lacks information regarding diagnosis cost and speed.

Further review on traditional laboratory testing approaches shows that, according to[(14)], Reverse Transcription Polymerase Chain Reaction (RT-PCR) features higher sensitivity and accuracy compared to other approaches. The literature reviewed reports 100% sensitivity and specificity and approximately 95.45% accuracy for RT-PCR, with an assay time ranging from 3 to 5 minutes. The Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)-based method, as reported in[(14)], achieves an average accuracy of 96.5% and a specificity of 98%, with an average assay time of 60 minutes.

Reverse Transcription Loop-Mediated Isothermal Amplification (RT-LAMP), Reverse Transcription Recombinase Polymerase Amplification (RT-RPA), and RAA methods, on the other hand, have an average accuracy of 96% and an assay time of 45 minutes. X-ray and lung ultrasound scans achieve 93% accuracy. In the study conducted by Hou et al.,[15], further experiments on developing CRISPR-based diagnostics for COVID-19 achieved up to 100% sensitivity, with an assay time of 40 minutes and material costs ranging from $0.7 to $3.5. Similarly, Kim et al.[16] experimented with COVID-19 testing costs using both simple and matrix pooling approaches, they estimated the cost per test as $70.

This paper seeks to close the existing research gap by implementing ML approaches on radiographic images and using traditional clinical testing methods such as Reverse Transcription Polymerase Chain Reaction (RT-PCR) as the reference for comparison to establish how the newly proposed approaches improve COVID diagnosis. The following research question was addressed:

How does the performance of the Machine Learning-Based Prediction diagnosis compare to traditional laboratory-based testing methods (e.g., RT-PCR) in terms of accuracy, precision, sensitivity, specificity, execution time, and cost?

2 Materials and Methods

2.1 Experimental settings

This study utilized a secondary radiography dataset, retrieved from the Kaggle data repository. The original dataset was compiled from publicly available posterior-to-anterior (PA) chest X-rays by a group of research collaborators from Qatar and Bangladesh. The experimental setting involved a placebo type of design with 3 study groups, on the final release these included; A control group with 10192 x-ray images from normal people, an experimental group with 3616 x-ray images from COVID-infected persons, and a group with viral pneumonia infected persons with 1345 images, and 6012 lung opacity images[17]. For this study, our focus was solely on distinguishing COVID-infected individuals from those with normal chest X-rays. Therefore, we disregarded the other categories. The data was divided into a training set (80%) and a testing set (20%) during the pre-processing step. The testing set was held out to evaluate the performance of the models on new data. Table 1 shows image breakdowns across training and testing splits. Figure 1 is a sample of 16 images of COVID-infected and Normal people chosen randomly from this dataset.

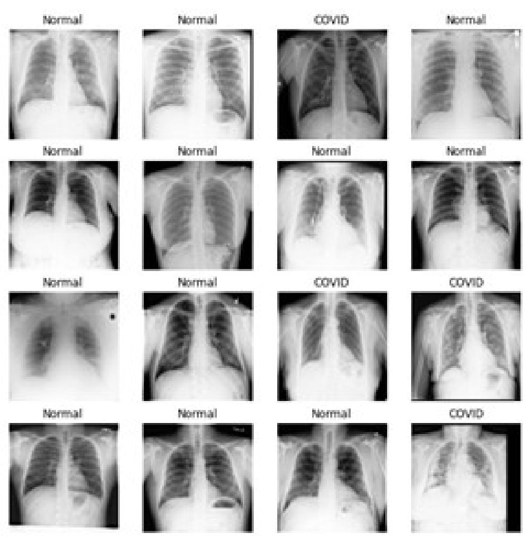

Fig 1. A Sample of 16 images, from data showing COVID and Normal images

2.2 Proposed approach

The proposed approach consists of two major phases; a feature extraction phase aimed at extracting relevant features for discriminating between COVID-infected chest images and normal chest images from the pixel data, and a classification phase that learns to distinguish between COVID-infected and normal chest images. Convolution Neural Networks (CNN or Convnet) have experienced a surge in popularity for visual imagery and video processing in recent years [18]. This is due to its architecture which allows zooming into images to get pixel information and pooling ability. Artificial Neural network (ANN) on the other hand uses weighting and activation functions to perform regression/classification tasks. Visual Geometry Group 16 (VGG16) is a popular architecture which combines CNN layer, ANN layers, ANN layers, and other convenience layer types such as pooling and dropout layers to perform Feature extraction and classification from Images. The architecture includes 13 CNN layers which extract features from 299 X299 images and 3 Dense ANN which perform classification. Figure 2 visualizes this representation.

During model training, regularization was performed using drop-out layers which randomly drop some connections to reduce the chances of overfitting, additionally the number of hidden layers was kept low to safeguard the model from overfitting. An early stopping callback overfitting signs.

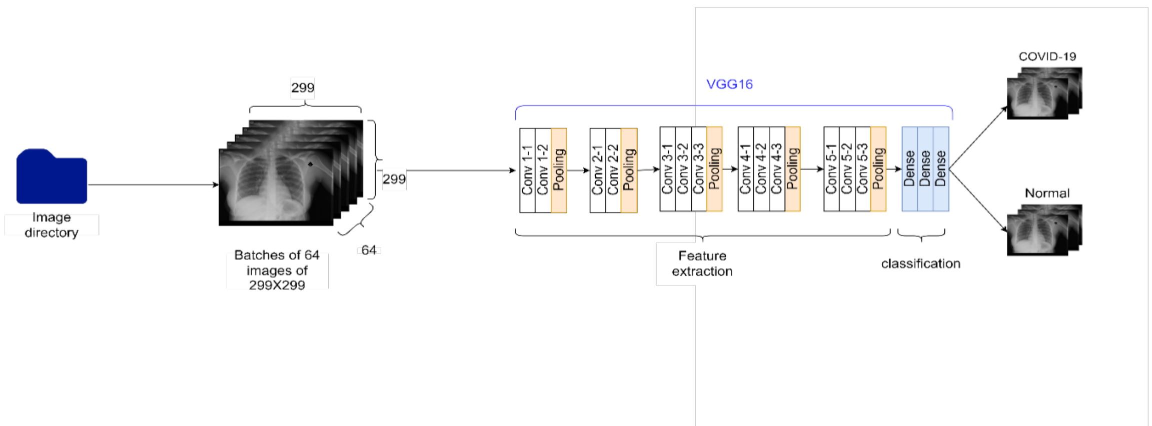

Fig 2. Visual representation of the proposed architecture

2.3 Benchmarking

The proposed approach was also tested on other real-life datasets of varied size and difficulty. The objective was to stress the model to see how it would perform in different real-life situations. This involves the Corona hack, a dataset chosen to assess the performance of this approach in the presence of competing outcomes and with low disease incidence, Chest X-Ray Scans which featured infants (0-5 years), a demographic more susceptible to respiratory-related mortalities [19], Patch Camelyon dataset which is a large and balanced dataset, and COVID19+PNEUMONIA+NORMAL Chest X-Ray Images dataset which is a well-balanced small dataset. Additionally, the Patch Camelyon dataset, characterized by its large and balanced nature, and the COVID19+PNEUMONIA+NORMAL Chest X-Ray Images dataset, which is well-balanced yet smaller in size. The proposed Artificial Neural Network (ANN) classifier will be compared against four other commonly used machines learning classifiers: Support Vector Machine (SVM), Logistic Regression (LR), Decision Tree (DT), and Random Forest Classifier (RF).

3 Results and Discussion

3.1 Proposed model

Generally, the model could correctly classify 98.41% of the testing images. Among those infected with the disease on the testing set, 97.54% were correctly identified. Conversely, 98.72% of the normal cases on the testing case were correctly classified. On the other hand, of those images on the testing set, the model classified as having come from COVID patients 96.49% were truly from COVID cases. The F1 score for the model is 98.41%. Finally, the execution time used by the model to classify a single case was computed as 0.0039 seconds. Table 1 reports the performance metrics for the model. Investigation of the model fitting history shows volatile values for accuracy/loss on the testing set, during the initial epochs, this improves gradually to stabilize around 60 epochs. Generally, we see that accuracy, as well as loss values on the testing set, doesn't differ largely from what we got from the training set. This means that the model doesn't suffer severely from overfitting. Conversely, accuracy values on both the testing and training sets are high, meaning that the model was able to learn enough from the dataset (no underfitting). Similarly, specificity and sensitivity remain tight together, indicating a good balance in performance. Figure 3 visualizes model training loss and accuracy during training.

Table 1. Image breakdown across splits of COVID-19 Radiography database

	Overall data	Training set	Testing set
Normal Images	10192	8163	2029
COVID images	3616	2884	732
Total images	13808	11047(80%)	2761(20%)
Number of batches	433	346	87

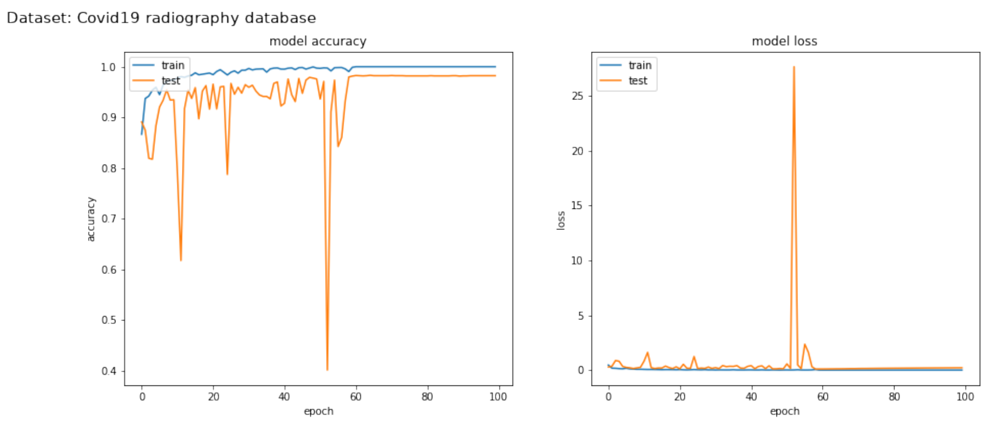

Fig 3. A visualization of the trend in accuracy and loss of the proposed model during training

3.2 Benchmark results

Exposing the proposed model on the Corona hack dataset showed that exposing the proposed approach to an unbalanced dataset, with limited COVID images and sufficient images from other outcomes leads to underfitting as well as low specificity compared to sensitivity. Similarly, the results of benchmarking the approach on the COVID-19+PNEUMONIA+NORMAL Chest X-Ray Images dataset show that the proposed approach works well with balanced outcome groups, specificity and sensitivity are well-balanced in such an environment. Benchmarking the approach on Chest X-ray scans (Pneumonia) data yielded low specificity and accuracy in general, the model tends to overfit on these kinds of images. On the patch camelyon dataset, we see the model, fits slightly well on huge datasets, yielding fairly high accuracy and sensitivity. Table 2 reports the performances on different datasets.

Table 2. Performance of our approach in varied datasets

Measure	Main dataset	Corona hack data	Chest X-Ray Scans	COVID+Pneumonia+Normal data	Patch camelyon data
Accuracy	**0.9841**	0.6075	0.7147	0.9985	0.8431
Sensitivity	**0.9754**	0.7500	0.9974	1.0000	0.9128
Specificity	**0.9872**	0.2009	0.2436	0.9973	0.7735
Precision	**0.9649**	0.7500	0.6873	0.9968	0.8013
F1-score	**0.9841**	0.6075	0.7147	0.9985	0.8431
Execution time	0.0039	0.0039	0.0039	0.0049	0.0006

By comparing the proposed model to other common machine learning approaches we see that the suggested approach was the best in regard to accuracy, followed by the support vector machine. It although capitalizes on specificity to improve accuracy, this means that the model correctly detects normal cases than the rest of the approaches, it's slightly lower higher false positives will frequently cause false COVID alarms compared to the other models. Conversely, the other models have high sensitivity meaning they detect the disease better. Table 3 reports the performances of different classifiers compared to the proposed ANN classifier

Table 3. Comparison of model performance

Measure	Main Model	Logistic regression	Decision tree	SVM	Random forest
Accuracy	**0.9841**	0.9797	0.9018	0.9812	0.9095
Sensitivity	0.9754	0.9867	0.9374	**0.9872**	0.9078
Specitivity	**0.9872**	0.9604	0.8033	0.9645	0.9139
Precision	0.9649	0.9857	0.9296	**0.9872**	0.9669
F1_score	0.9841	0.9862	0.9335	**0.9872**	0.9365
Excecution time	0.0074	0.0005	**<0.0001**	0.0924	12.394

The proposed approach in this study has demonstrated superior performance compared to previous frameworks and models. Specifically, it outperformed [6] COVID-CAPS framework, achieving an accuracy of 95.7% and a specificity of 95.80%. It also outperformed the EfficientNet-B0 and curvelet transform EfficientNet-B0 approaches by [8], which achieved lower accuracy results. Kumar et al. [12] recommended ResNet18 as the fastest and most accurate, yet the VGG16 model proposed in this study achieved relatively higher accuracy. On the other hand, [3] focused on the automatic prediction of COVID-19 and found it challenging to determine the relationship between COVID-19 and pneumonia. Their approach with the XGboost classifier yielded lower accuracy compared to the current study. Öztürk et al., [20] achieved slightly lower accuracy with a Darknet classifier, while [11] achieved similar results with the KE Sieve Neutral Network. Johri et al. [21] suggested the use of transfer learning with tInceptionV3 and VGG16 models, incorporating SVM to improve accuracy, and achieved an accuracy of 98.34%. In terms of execution time and cost, the model in this study outperformed previous studies reported by [14]. Specifically, in terms of assay time, the testing costs reported by [16] were similar to the costs of taking X-ray scans reported by [22].

4 Conclusion

Overall, this study successfully implemented machine learning classifiers for detecting COVID-19 infections using x-ray images. The classifiers achieved high accuracy (98.41%) and a balanced prediction between normal (sensitivity=97.54) and COVID-19 cases (specificity=98.52%). The approach combines CNN feature extraction and SVM proved to be effective, particularly when the misclassification cost for COVID-19 cases was relatively high. The study emphasizes the importance of large and well-balanced datasets in improving model fitting with fewer training epochs.

The introduction of the AI approach also contributes to cost efficiency by offering a cheaper screening and diagnostic system that only requires chest X-ray images, eliminating the need for expensive medical testing kits and laboratory equipment. The system can be implemented outside of a laboratory setting. This study provides valuable literature and successful attempts at image classification, highlighting the significance of integrating AI technologies in patient classification, diagnosis, screening, and testing. Future researchers can refer to this study, draw evidence from its findings, and build upon its results to further improve their research.

The proposed approach relies on chest X-ray scans to predict the presence of COVID. Conversely, researchers such as [23] and [24] have provided evidence that exposure to X-ray scans is a risk factor for cancer. This means that the proposed approach is limited by the fact that patients can only be screened this way just a few times. Moreover, patients who had prior exposures to such scans might not be eligible for screening using these new approaches.

If adopted, the plan is thought to be extremely beneficial for everyone and will have a significant impact on decreasing the COVID rate. The COVID-19 testing approach also achieved a significant milestone as a result. To raise accuracy and provide more models to compare deep learning models to, we will need to enhance CNN architecture in the future. Additionally, by encouraging users to keep their login information, and data, and open their comment areas for greater improvement, we can scale up the web application. Misclassification costs are too high for false negative classifications as they put populations at risk of discovering their COVID status too late for this reason, the focus needs to shift from overall accuracy to sensitivity, on the other hand, SVM classifier achieved significantly better performance based on sensitivity and F1 score, which means future literature trying SVM, classifier and its accuracy is needed.

References

1) Pinter G, Felde I, Mosavi A, Ghamisi P, Gloaguen R. COVID-19 Pandemic Prediction for Hungary; A Hybrid Machine Learning Approach. *Mathematics*. 2020;8(6):890–890. Available from: https://doi.org/10.3390/math8060890.
2) Khan HR, Hossain A. Machine Learning Approaches Reveal That the Number of Tests Do Not Matter to the Prediction of Global Confirmed COVID-19 Cases. *Frontiers in artificial intelligence*. 2020;3. Available from: https://doi.org/10.3389/frai.2020.561801.
3) Mehta M, Julaiti J, Griffin P, Kumara S. Early Stage Machine Learning–Based Prediction of US County Vulnerability to the COVID-19 Pandemic: Machine Learning Approach. *JMIR Public Health and Surveillance*. 2020;6(3). Available from: https://doi.org/10.2196/19446.
4) Alballa N, Al-Turaiki I. Machine learning approaches in COVID-19 diagnosis, mortality, and severity risk prediction: A review. *Informatics in Medicine Unlocked*. 2021;24:100564–100564. Available from: https://dx.doi.org/10.1016/j.imu.2021.100564.
5) Akhter Y, Singh R, Vatsa M. AI-based radiodiagnosis using chest X-rays: A review. *Frontiers in big data*. 2023;6. Available from: https://doi.org/10.3389/fdata.2023.1120989.
6) Zoabi Y, Deri-Rozov S, Shomron N. Machine learning-based prediction of COVID-19 diagnosis based on symptoms. *npj digital medicine*. 2021;4. Available from: https://www.nature.com/articles/s41746-020-00372-6.pdf.
7) Saygili A. A new approach for computer-aided detection of coronavirus (COVID-19) from CT and X-ray images using machine learning methods. *Applied Soft Computing*. 2021;105:107323–107323. Available from: https://doi.org/10.1016/j.asoc.2021.107323.
8) Altan A, Karasu S. Recognition of COVID-19 disease from X-ray images by hybrid model consisting of 2D curvelet transform, chaotic salp swarm algorithm and deep learning technique. *Chaos Solitons & Fractals*. 2020;140:110071–110071. Available from: https://doi.org/10.1016/j.chaos.2020.110071.
9) Kumar M, Shakya D, Kurup V, Suksatan W. COVID-19 prediction through X-ray images using transfer learning-based hybrid deep learning approach. *Materials Today: Proceedings [Internet*. 2021. Available from: https://doi.org/10.1016/j.matpr.2021.12.123.
10) Huang X, Stubblefield J. COVID19 diagnosis using chest x-rays and transfer learning.. . Available from: https://doi.org/10.1101/2022.10.09.22280877.
11) Abbas A, Abdelsamea MM, Gaber MM. Classification of COVID-19 in chest X-ray images using DeTraC deep convolutional neural network. *Applied Intelligence*. 2021;51:854–64. Available from: https://doi.org/10.1007/s10489-020-01829-7.
12) Kumar R, Arora R, Reddy BM, Sahayasheela VJ, Buckchash H, Imran J. Accurate Prediction of COVID-19 using Chest X-Ray Images through Deep Feature Learning model with SMOTE and Machine Learning Classifiers. . 2020. Available from: https://doi.org/10.1101/2020.04.13.20063461.
13) Fu S, Fu X, Song Y, Minteer SD, Pan P, Tang T. Virologic and clinical characteristics for prognosis of severe COVID-19: a retrospective observational study in Wuhan, China. 2020. Available from: https://doi.org/10.1101/2020.04.03.20051763.
14) Filchakova O, Dossym D, Ilyas A, Tk K, Abdizhamil A, Bukasov R. Review of COVID-19 testing and diagnostic methods. *Talanta*. 2022;244:123409–123409. Available from: https://doi.org/10.1016/j.talanta.2022.123409.
15) Hou T, Zeng W, Yang M, Chen W, Ren L, Ai JW. Development and evaluation of a rapid CRISPR-based diagnostic for COVID-19. *PLOS Pathogens*. 2020;16(8):1008705–1008705. Available from: https://doi.org/10.1371/journal.ppat.1008705.
16) Kim EK, Kim JY, Sung H, Jo MW. Cost analysis of coronavirus disease 2019 test strategies using pooled reverse transcriptase-polymerase chain reaction technique. *Journal of Clinical Laboratory Analysis*. 2022. Available from: https://onlinelibrary.wiley.com/doi/pdfdirect/10.1002/jcla.24413.
17) Choudhary DK, Tripathi M, Shankar R. Reliability, availability and maintainability analysis of a cement plant: a case study. *International Journal of Quality & Reliability Management*. 2019;36(3):298–313. Available from: https://doi.org/10.1108/ijqrm-10-2017-0215.
18) Afzal S, Ghani SA, Hittawe MM, Rashid SF, Knio OM, Hadwiger M. Visualization and Visual Analytics Approaches for image and video datasets: A survey. . *ACM transactions on interactive intelligent systems*. 2023;13:1–41. Available from: https://doi.org/10.1145/3576935.
19) Tang YW, Sussman M, Liu D, Poxton I, Schwartzman J. Molecular Medical Microbiology – The Expanding Concept. In: and others, editor. Molecular Medical Microbiology. 2014. Available from: https://doi.org/10.1016/b978-0-12-397169-2.00001-9.
20) Öztürk T, Talo M, Yildirim EA, Baloğlu UB, Yıldırım Ö, Acharya UR. Automated detection of COVID-19 cases using deep neural networks with X-ray images. *Computers in Biology and Medicine*. 2020;121:103792–103792. Available from: https://doi.org/10.1016/j.compbiomed.2020.103792.
21) Johri S, Goyal M, SJain, Baranwal M, Kumar V, Upadhyay R. A novel machine learning-based analytical framework for automatic detection of COVID-19 using chest X-ray images. *International Journal of Imaging Systems and Technology*. 2021;31(3):1105–1119. Available from: https://dx.doi.org/10.1002/ima.22613. doi:10.1002/ima.22613.
22) Kalhor R, Norouzi S, Kiaei MZ, Emami M, Maleki YS, Ramandi FF. The Calculation of the Costs of CT-scan Services at Shahid Rajaei Hospital in Qazvin, Iran Using the ABC Model and Comparison with the Tariff of the Iranian Ministry of Health. *International journal of health and life sciences*. 2020;6. Available from: https://doi.org/10.5812/ijhls.100901.
23) Schultz CH, Fairley R, Murphy LS, Doss M. The Risk of Cancer from CT Scans and Other Sources of Low-Dose Radiation: A Critical Appraisal of Methodologic Quality. *Prehospital and Disaster Medicine*. 2020;35(1):3–16. Available from: https://dx.doi.org/10.1017/s1049023x1900520x. doi:10.1017/s1049023x1900520x.
24) Shao YH, Tsai K, Kim S, Wu YJ, Demissie K. Exposure to Tomographic Scans and Cancer Risks. *JNCI Cancer Spectrum*. 2020;4(1). Available from: https://dx.doi.org/10.1093/jncics/pkz072.