EMITTER International Journal of Engineering Technology
Vol. 10, No. 1, June 2022, pp. 183~194
DOI : 10.24003/emitter.v10i1.672

A Machine learning Classification approach for detection of Covid 19 using CT images

Suguna G C[1], Veerabhadrappa S T[1]*, Tejas A[2], Vaishnavi P[2], Sudarshan E[2], Raghunandan V Gowda[2], Panahami R Udupa[2], Spoorthy R[2], Smitha Reddy[2]

[1] Department of Electronics and Communication Engineering,
JSS Academy of Technical Education, Bengaluru, Karnataka, India
[2] Undergraduate students, Department of Electronics and Communication Engineering, JSS Academy of Technical Education, Bengaluru, Karnataka, India
Correspondence Author: veerabhadrappast@jssateb.ac.in

Received February 10, 2022; Revised March 14, 2022; Accepted April 18, 2022

ABSTRACT:

Coronavirus disease 2019 popularly known as COVID 19 was first found in Wuhan, China in December 2019. World Health Organization declared Covid 19 as a transmission disease. The symptoms were cough, loss of taste, fever, tiredness, respiratory problem. These symptoms were likely to show within 11 –14 days. The RT-PCR and rapid antigen biochemical tests were done for the detection of COVID 19. In addition to biochemical tests, X-Ray and Computed Tomography (CT) images are used for the minute details of the severity of the disease. To enhance efficiency and accuracy of analysis/detection of COVID images and to reduce of doctors' time for analysis could be addressed through Artificial Intelligence. The dataset from Kaggle was utilized to analyze. The statistical and GLCM features were extracted from CT images for the classification of COVID and NON-COVID instances in this study. CT images were used to extract statistical and GLCM features for categorization. In the proposed/prototype model, we achieved the classification accuracy of 91%, and 94.5% using SVM and Random Forest respectively.

***Keywords:* -** Covid*, SVM, Random Forest, Computed Tomography, GLCM*

1. INTRODUCTION

Since January 2020, the novel coronavirus (nCoV) infection has spread throughout the world. Specific COVID-19 drugs are unavailable, it is critical to investigation the disease at an early stage and isolate afflicted patients immediately. The comparatively gradual development of symptoms, which allows for extensive transmission by asymptomatic carriers, contributes to the high rates of infection [1]. With today's travel society's global connectivity, this virus quickly spread over the world [1], resulting in a pandemic [2, 3]. The RT-PCR and rapid antigen biochemical tests were done for the detection of COVID 19. In addition to biochemical tests, X-Ray and

Computed Tomography (CT) images are used for the minute details of the severity of the disease. RT-PCR tests were used in the majority, but due to the delay of RT-PCR reports physicians suggested lung X-rays and CT scans. Research groups reported that CT scan images indicate lung parenchymal damage, which consists of great consolidation and interstitial inflammation in covid patients [4,5].

CT and chest X-rays are excellent tools for imaging the lung in COVID-19 infections. Unlike swab tests, CT and X-rays disclose the spatial location of probable pathology and the level of damage [6]. CT Imaging provides the advantage of being very sensitive, fast turnaround, and visualizing the degree of lung infection. Imaging's drawback is that it has a low specificity, making it difficult to distinguish between different types of lung infection, especially when the infection is severe. However, a CT scan does not provide information about the variant which is causing the infection. During the widespread of covid 19, radiologists were highly occupied, and could not read the reports timely [7]. To overcome this situation, many research groups have developed machine learning, deep learning techniques to classify as covid or non-covid [8-11]. The machine-learning (ML) feature-based methods have a wide range of applications, including many biomedical signal and image processing applications such as arrhythmia classification, epileptic seizure, and cancer detection [12-19]. Radiologists can benefit from computer-aided diagnostic systems to improve diagnostic accuracy. Researchers are currently employing learning features based on the texture, shape, and morphological characteristics of lung detection.

2. RELATED WORK

An automated detection method is essentially required to aid in the screening of COVID-19 pneumonia using chest CT imaging. Professional medical experts must manually analyze chest CT images, which is a time-consuming and resource-intensive operation. Many researchers have attempted to analyze CT image datasets with various machine learning and deep learning techniques [20,21,22]. Segmentation, Machine Learning, and Deep Learning were the most commonly used approaches. Image processing techniques were used to extract the lung section, equalise histograms using a transformation created by the intensities of the infected areas, and boost image contrast.

M. Barstugan et al. proposed lung region and lung lesion oriented segmentation methods in Sars Covid 19. The Lung region-oriented process deals with lobes and whole parts of the lung from the background of CT images. The Lung lesion process deals with separate lesions in the lung and from lung regions as the lesions or nodules are small compared to rest regions. The U-Net architecture has been used for segmentation with symmetric decoding and encoding signals [23,24].

N. Yang et al [25] proposed a support vector machine to classify COVID-19 and other types of pneumonia with an accuracy of 89.83%. COVID-19 patients were classified using a combination of the Gray Level Co-occurrence Matrix and the SVM model [24]. To diagnose the covid cases, self-supervised features were extracted and deep learning techniques were used. A two-stage Convolutional Neural Network-based classification was proposed for detecting COVID-19 using chest computed tomography scans.

Computational imaging techniques based on deep learning appear to be useful for assessing positive COVID-19 cases [26]. Several research groups proposed adaptive features of CT scan images used to classify covid cases with the support of deep learning methods [9,10,11]. Lal Hussain et.al, proposed the texture-based features with machine learning techniques to increase the accuracy of the classifier [27]. This paper includes various statically and GLCM features are used to classify the Covid and Non-Covid Images using SVM and Random Forest Classifier.

3. **ORIGINALITY**

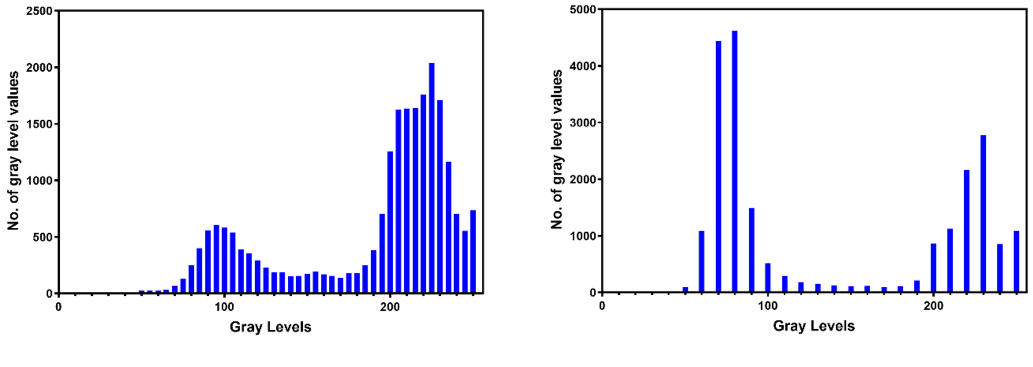
The methodology consists of four sections which include pre-processing, feature extraction, feature selection, and classification. In this study, the dataset from Kaggle was used for the analysis and classification of covid and non-covid cases [28,29].

3.1 Pre-Processing

A Kaggle database consisting of 1252 and 1229 CT scan images related to covid and non-covid cases respectively with the description of the clinical findings. The images were equally selected in number for the processing purpose and were balanced. The images are of their original and regular size, which indeed has been pre-processed and modeled. In the given data set the intensity ranges and contrast must be similar across the database, therefore, all obtained CT images were resized to 150 x150 pixels, and all those were converted into grayscale.

3.2 Feature extraction

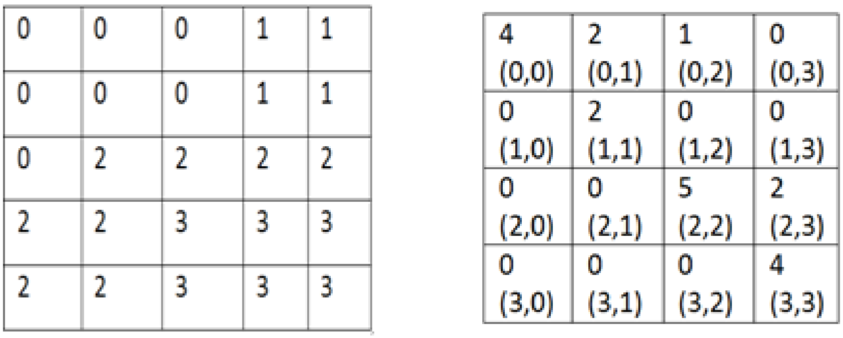
Feature extraction is the process of extracting features from raw data using domain knowledge. Predictive models use features to influence results. Statistical and grey level co-occurrence matrix (GLCM) features are extracted as significant categories of features for feature engineering. Statistical features, as shown in Figure 1, deal with the appearance on a grey level scale based on a histogram. The first order to fourth-order statistical features was extracted for the classification.



(a) (b)

Figure 1. Histogram for (a) covid and (b) non-covid CT-scan images

In addition to statistical features, GLCM features are obtained from the gray level co-occurrence matrix as shown in Figure 2. The GLCM features exhibit the relationship between the pixel values in different angles and different distances. These statistics are functions of distance and orientation. The GLCM features such as correlation, contrast, angular second moment, difference entropy, and difference variance were extracted.



0	0	0	1	1
0	0	0	1	1
0	2	2	2	2
2	2	3	3	3
2	2	3	3	3

4 (0,0)	2 (0,1)	1 (0,2)	0 (0,3)
0 (1,0)	2 (1,1)	0 (1,2)	0 (1,3)
0 (2,0)	0 (2,1)	5 (2,2)	2 (2,3)
0 (3,0)	0 (3,1)	0 (3,2)	4 (3,3)

(a) (b)

Figure 2. Gray level co-occurrence matrix

The Pearson correlation coefficient quantifies the degree of linear correlation between two sets of data., which is used for the selection of features based on highly correlated independent features. Figure 3 shows the relationship between the correlation coefficient and their indications. The independent features with more than 85% positively correlated were removed.

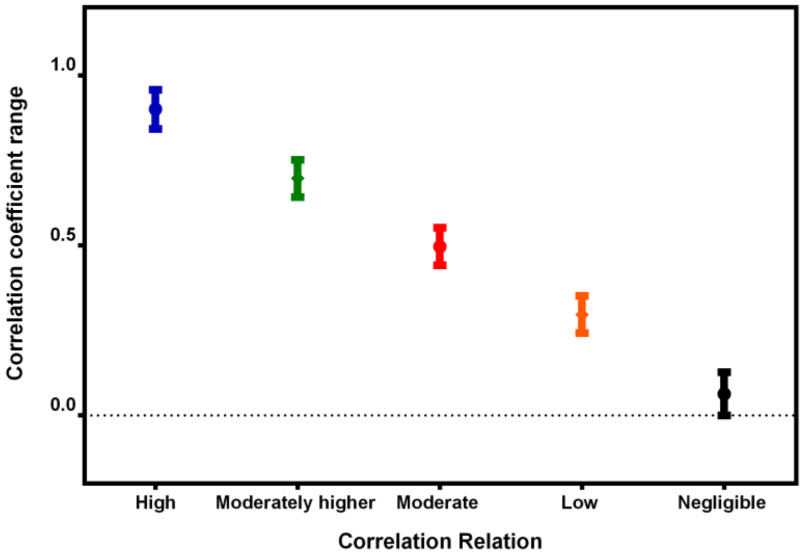


Figure 3. Correlation coefficient and their indications

4. **SYSTEM DESIGN**

After feature engineering and feature selection, classification is done using Random Forest Classifiers (RFC) and Support Vector Machine (SVM) classifiers. RFC and SVM are robust algorithms and could accommodate a large dataset.

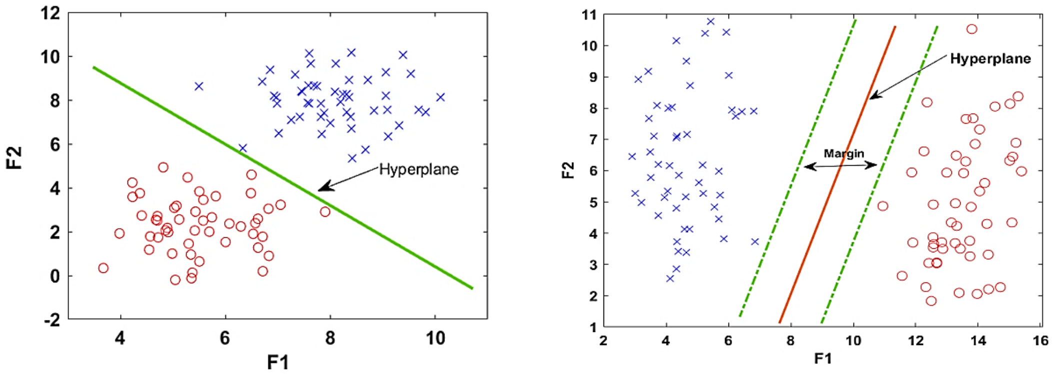
4.1 Support Vector Machine (SVM)

The Support Vector Machine (SVM) is a supervised machine learning tool that can be used to solve problems like classification and regression.The classifier uses the hyperplane with the most margin to separate data points. SVM mapping the feature vectors to higher-dimensional space using a nonlinear method and further classifying using linear classification method. The support vectors could be used to maximize the classification margin. The data points, support vectors, and hyperplane are shown in Figure 4.

The SVM algorithm attempts to maximize the difference between the data points and the hyperplane. Hinge loss is a loss function that aids in margin maximization.

$$C(F1,F2,g(x)) = \begin{cases} 0 & F2 * g(x) \geq 1 \\ 1 - F2 * g(x) & elsewhere \end{cases} \quad \text{------ (1)}$$

The cost function C will be zero if both predicted and actual values have the same sign. The cost function now includes a regularisation parameter. The goal of the regularisation parameter is to strike a balance between margin maximization and loss.



(*a*) (b)

Figure 4. SVM Classification features with Hyperplane (a) without margin (b) with margin

4.2 Random Forest Classifier (RFC)

Random forest was an ensemble supervised machine learning algorithm first introduced by Breiman [30,31]. Which is a combination of more than one algorithm of the same or different kinds for classifying objects. The number of decision trees in the model depends on the dataset and is tuned using the hyperparameter 'n_estimators'. The random forest model combines the predictions made by the estimators which produce a more optimal solution. Random Forest algorithm is a combination of Bagging and Decision Tree algorithm. The random forest splits the nodes into sub-nodes randomly and selects the best features out of these sub-nodes results in a better classifier. A random forest uses an averaging technique to improve predictive accuracy and to reduce over-fitting. Averaging means to take an average value from the output produced by the 'n' number of decision trees. It also uses majority voting at times depending on the scenario. In classification tasks, majority voting is preferred while in the case of regression tasks, the generally averaging process is carried out. This process of taking an average or majority voting is called aggregation. Random forest, decision trees are combined in parallel which results in lower variance as compared to outputs of individual decision trees as the output depends on multiple decision trees. Random forest is good at handling high dimensionality and heterogeneous feature types. The most important hyperparameter inside the random forest algorithm includes 'n_estimators', 'max_depth' and 'min_samples_split'. The n_estimators refer to the number of decision trees used in the random forest algorithm defined by the Gini index, non-parametric measurement of classifier or regressor, and; maximum depth denotes the height of the tree. Gini index (GI) is represented with equation (2). The deeper the tree is the more split it has and hence captures more information about the data. An optimum value is to be set for maximum depth as it overfits for larger values. Minimum samples split is the number of samples, which are required to split an internal node.

$$GI(X) = 1 - \sum_{k=1}^{n_{estimators}} x_j^2 \qquad \text{------ (2)}$$

where x_j is relative frequency of class j , X=$\{x_1, x_2\ x_3 \ldots\ldots\}$

The decision tree will be built based *m* different features. Prediction from each tree will be stored and the best solution is selected through voting.

5. EXPERIMENT AND ANALYSIS

CT scan images of covid and non-covid patients are shown in Figure 5. In this study, the SVM and random forest are used for the classification of covid and non-covid. Further, the data was split into two categories such as training and testing in the ratio of 80:20 percent respectively. The features are extracted based on the histogram of the image and the co-occurrence matrix. As seen in Figure 5 the histogram of covid and non-covid the spread of gray level varies for covid its more concentrated on the white region that near to 255 values as the covid images contain white patches in the infected regions. The non-covid images show more concentration of darker regions than near to 0 values.

The statistical and GLCM features vary for covid and non-covid images which give highly correlated features for classification. The significance of the covid and non-covid features was determined using the t-test and ANOVA test. The statistical and GLCM features exhibit the most significant between the covid and non-covid with a p-value less than 0.0001. The mean and standard deviation of the statistical and GLCM features are depicted in Table 1. Table 2 indicates the classification results of statistical and GLCM features for covid versus non-covid using two different classifiers. The data contains balanced data with both classes being almost equal. Further, the features selection was based on the Pearson correlation which detects the highly correlated independent features. The independent features with more than 85% correlated were removed for the process of classification.

K-fold cross-validation was used to improve the classifiers' performance. The data is randomly split into training and testing data, and SVM or Random Forest model is fitted to the training set and evaluated on the test set in the cross-validation process. In each iteration model, chooses a different set of training and test data. In this study, two trials of cross-validation 5-fold and 10-fold were performed on the dataset. The mean and standard deviation of 5-fold and 10-fold cross-validation is depicted in Table 2. The cross-validation avoids the problem of overfitting the model and enhances the accuracy.

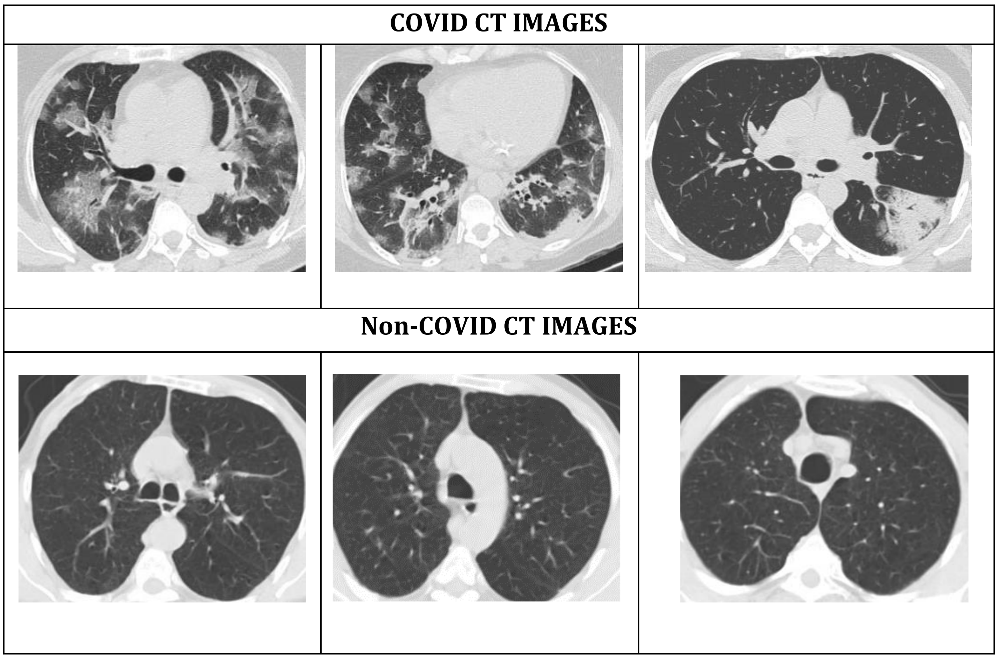


Figure 5. Sample of Covid 19 and Non- Covid images from Kaggle database

Table1. Features extracted from covid and non- covid for the classification

	Parameters	Covid		Non-Covid	
Statistical Features		**Mean**	**Std. Deviation**	**Mean**	**Std. Deviation**
	image_mean	168.1000	20.2200	158.5000	12.6700
	image_std	63.7800	9.5000	69.6400	4.4220
	img_var	4158.000	1115.0000	4869.000	620.3000
	image_skew	-0.3815	0.5784	0.0170	0.3086
	image_kurtosis	-1.1800	1.4520	-1.7410	0.2783
	1std entropy	14.3300	0.0502	14.3100	0.0324
	10th percentile	54.9900	28.5900	57.7500	25.3200
	90th percentile	84.3900	19.3400	80.5400	6.5390
	median	181.5000	42.5600	154.4000	49.8900
	Mean Absolute Deviation (MAD)	119.8000	37.1100	140.8000	12.5400
	Median Absolute Deviation (Med AD)	181.5000	42.5600	154.4000	49.8900
	range	200.6000	11.4600	201.5000	10.2000
	RmeanAD	59.1400	12.2900	67.0400	5.1200
	IQR	235.4000	7.8420	239.2000	7.6750

GLCM Features	Angular Second Moment	0.0024	0.0079	0.0036	0.0110
	Contrast	466.0000	189.4000	459.2000	205.4000
	Correlation	0.9431	0.0217	0.9536	0.0176
	Difference Entropy	4.7490	0.4006	4.5870	0.3234
	Difference Variance	0.0002	0.0001	0.0003	0.0001
	dissimilarity	11.1600	2.6690	10.3400	2.3770
	energy	0.0375	0.0308	0.0424	0.0419
	Entropy	11.4500	0.7584	11.1300	0.6635
	Homogeneity	0.1846	0.0714	0.2083	0.0607
	Information Measure of Correlation 1	-0.2734	0.0473	-0.2859	0.0393
	Information Measure of Correlation 2	0.9838	0.0097	0.9857	0.0081
	Inverse Difference Moment	0.1846	0.0714	0.2083	0.0607
	Maximal Correlation Coefficient	4.6190	0.4267	4.3060	0.4151
	Sum Average	335.8000	40.7000	316.5000	25.4300
	Sum Entropy	7.5740	0.3734	7.4390	0.3609
	Sum of Squares: Variance	4160.000	1113.0000	4868.000	617.3000
	Sum Variance	16175.00	4339.0000	19012.00	2348.0000

Table 2. Performance metrics of the Classifiers

Classifier	**Validation**	**Sensitivity**	**Specificity**	**Accuracy**	**Precision**	**F1 Score**
SVM	1-fold	93.43	86.7	90.1	88.2	83.0
	5-fold	94.29±2.06	86.92 ± 2.0	90.56±1.50	87.61±2.08	83.2±2.709
	10-fold	94.42±2.63	88.06±3.79	91.13±1.99	88.52±4.05	84.05 ±3.59
Random Forest	1-fold	95.62	94.02	94.8	94.4	90.5
	5-fold	95.65±2.12	96.41±0.65	96.04±1.29	96.29±0.833	92.27 ±2.61
	10-fold	96.07±1.86	96.43±1.97	96.25±1.12	96.31±2.05	92.64 ±2.38

6. CONCLUSION

Appropriate diagnosis of Covid-19 positive patients was extremely important for providing them in time and also for preventing other people from getting a coronavirus infection. The machine learning approach such as random forest and SVM classifiers were employed in this study to diagnose the Covid-19 patient, and the findings showed that radiologists can readily observe and classify the patient as Covid-19 positive or negative using the aforementioned approach. The accuracy of the models could be improved by more precisely by normalizing the dataset. Finally, two ML classification models were built and it can be seen that the random forest classification algorithm outperforms the SVM in terms of accuracy in this dataset.

Acknowledgments

We thank all the members of the Artificial Intelligence (AI) Club, JSS Academy of Technical Education, Bengaluru for their useful suggestions during this work.

References

[1] Biscayart C, Angeleri P, Lloveras S, Chaves TSS, Schlagenhauf P, Rodríguez-Morales**, The next big threat to global health? 2019 novel coronavirus (2019-nCoV): What advice can we give to travellers? – Interim recommendations January 2020, from the Latin-American society for Travel Medicine (SLAMVI)**, Travel medicine and infectious disease Vol.33, pp. 101567, 2020.

[2] Carlos WG, Dela Cruz CS, Cao B, Pasnick S, Jamil S, **Novel Wuhan (2019-nCoV) coronavirus**, Am J Respir Crit Care Med, pp. P7-8, 2020.

[3] Munster VJ, Koopmans M, van Doremalen N, van Riel D, de Wit E, **A novel coronavirus emerging in China—key questions for impact assessment**, New England Journal of Medicine 382, Vol.no. 8, pp. 692-694, 2020.

[4] Chung M, Bernheim A, Mei X, Zhang N, Huang M, Zeng X, **CT imaging features of 2019 novel coronavirus (2019-nCoV)**, Radiology, Vol.No 295, pp.202–207, 2020.

[5] Fang Y, Zhang H, Xu Y, Xie J, Pang P, Ji W, **CT manifestations of two cases of 2019 novel coronavirus (2019-nCoV) pneumonia**, Radiology, Vol.No 295, pp.208–209, 2020.

[6] Wong HYF, Lam HYS, Fong AH-T, Leung ST, Chin TW-Y, Lo CSY, et al**, Frequency and distribution of chest radiographic fndings in COVID-19 positive patients**, Radiology, Vol.No 296, pp. E72-E78, 2020.

[7] Yang, X., He, X., Zhao, J., Zhang, Y., Zhang, S. and Xie, P., 2020, **COVID-CT-dataset: a CT scan dataset about COVID-19**, *arXiv preprint arXiv:2003.13865*, March 2020.

[8] Sun, Liang, Zhanhao Mo, Fuhua Yan, Liming Xia, Fei Shan, Zhongxiang Ding, Bin Song et al, **Adaptive feature selection guided deep forest**

for covid-19 classification with chest ct, *IEEE Journal of Biomedical and Health Informatics* 24, Vol.No. 10, pp. 2798-2805, 2020.

[9] Farooq, Junaid, and Mohammad Abid Bazaz, **A novel adaptive deep learning model of Covid-19 with focus on mortality reduction strategies**, *Chaos, Solitons & Fractals* Vol.No138 , pp.110148, 2020.

[10] Wang, Shuai, Bo Kang, Jinlu Ma, Xianjun Zeng, Mingming Xiao, Jia Guo, Mengjiao Cai et al, **A deep learning algorithm using CT images to screen for Corona Virus Disease (COVID-19)**, European radiology, pp.1-9, 2021.

[11] Wang, G., Liu, X., Li, C., Xu, Z., Ruan, J., Zhu, H., Meng, T., Li, K., Huang, N. and Zhang, S, **A noise-robust framework for automatic segmentation of COVID-19 pneumonia lesions from CT images**, IEEE Transactions on Medical Imaging, Vol.No 39, pp.2653-2663, 2020.

[12] Ebrahimi, Zahra, Mohammad Loni, Masoud Daneshtalab, and Arash Gharehbaghi, **A review on deep learning methods for ECG arrhythmia classification**, *Expert Systems with Applications: X* 7 pp. 100033, 2020.

[13] Jun, Tae Joon, Hoang Minh Nguyen, Daeyoun Kang, Dohyeun Kim, Daeyoung Kim, and Young-Hak Kim. **ECG arrhythmia classification using a 2-D convolutional neural network**. *arXiv preprint arXiv:1804.06812*, 2018.

[14] Nagabushanam, P., S. Thomas George, Praharsha Davu, P. Bincy, Meghana Naidu, and S. Radha. **Artifact Removal using Elliptic Filter and Classification using 1D-CNN for EEG signals**, In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, IEEE, pp. 551-556, 2020.

[15] Mirowski, Piotr, Deepak Madhavan, Yann LeCun, and Ruben Kuzniecky, **Classification of patterns of EEG synchronization for seizure prediction**, Clinical neurophysiology 120, Vol.No. 11, pp.1927-1940, 2009.

[16] Parmar C, Bakers FCH, Peters NHGM, Beets RGH, **Deep learning for fully-automated localization and segmentation of rectal cancer on multiparametric**, Sci Rep,Vol.No. 7, pp.1–9, 2017

[17] Oakden-Rayner L, Carneiro G, Bessen T, Nascimento JC, Bradley AP, Palmer LJ, **Precision radiology: predicting longevity using feature engineering and deep learning methods in a radiomics framework**, Sci Rep. Vol.No. 7, pp.1648, 2017.

[18] Cruz JA, Wishart DS, A**pplications of machine learning in cancer prediction and prognosis**, Cancer Inform, p.117693510600200030, Jan 2006.

[19] Doyle S, Hwang M, Shah K, Madabhushi A, Feldman M, Tomaszeweski J, **Automated grading of prostate cancer using architectural and textural image features**, 4th IEEE International Symposium on Biomedical Imaging From Nano to Macro. IEEE, pp.1284–7, 2007.

[20] Pathak, Yadunath, Prashant Kumar Shukla, Akhilesh Tiwari, Shalini Stalin, and Saurabh Singh, **Deep transfer learning based classification model for COVID-19 disease**, Irbm, May 2020.

[21] Li, Kunwei, Yijie Fang, Wenjuan Li, Cunxue Pan, Peixin Qin, Yinghua Zhong, Xueguo Liu, Mingqian Huang, Yuting Liao, and Shaolin Li, **CT image visual quantitative evaluation and clinical classification of coronavirus disease (COVID-19)**, *European radiology* 30, Vol.No. 8, pp. 4407-4416, 2020.

[22] Gilanie, G., Bajwa, U.I., Waraich, M.M., Asghar, M., Kousar, R., Kashif, A., Aslam, R.S., Qasim, M.M. and Rafique, H, **Coronavirus (COVID-19) detection from chest radiology images using convolutional neural networks**, Biomedical Signal Processing and Control, 66, p.102490, 2021.

[23] M. Barstugan, U. Ozkaya, and S. Ozturk, **Coronavirus (COVID-19) Classification using CT Images by Machine Learning Methods**, arXiv Prepr. arXiv2003.09424, no. 5, pp. 1–10, 2020

[24] Barstugan, M., Ozkaya, U. and Ozturk, S., 2020. **Coronavirus (covid-19) classification using ct images by machine learning methods**. *arXiv preprint arXiv:2003.09424*.

[25] N. Yang et al., **Diagnostic classification of coronavirus disease 2019 (COVID-19) and other pneumonias using radiomics features in CT chest images**, Artif. Intell. Mach. Learn., vol. 2019, pp. 1–11, 2020.

[26] Li, Y., Wei, D., Chen, J., Cao, S., Zhou, H., Zhu, Y., Wu, J., Lan, L., Sun, W., Qian, T. and Ma, K., **Efficient and effective training of covid-19 classification networks with self-supervised dual-track learning to rank**, *IEEE Journal of Biomedical and Health Informatics*, Vol.No. *24*(10), pp.2787-2797, 2020.

[27] Hussain, Lal, Tony Nguyen, Haifang Li, Adeel A. Abbasi, Kashif J. Lone, Zirun Zhao, Mahnoor Zaib, Anne Chen, and Tim Q. Duong, **Machine-learning classification of texture features of portable chest X-ray accurately classifies COVID-19 lung infection**, *BioMedical Engineering OnLine* 19, Vol.No. 1, pp. 1-18,2020.

[28] Sarker, S., Jamal, L., Ahmed, S.F. and Irtisam, N., 2021. **Robotics and artificial intelligence in healthcare during COVID-19 pandemic: A systematic review**. Robotics and autonomous systems, 146, p.103902.

[29] https://www.kaggle.com/plameneduardo/sarscov2-ctscan-dataset?select=COVID

[30] Breiman, Leo, **Random forests**, *Machine learning* 45,pp.5-32,2001.

[31] Sarica, A., Cerasa, A. and Quattrone, A., **Random forest algorithm for the classification of neuroimaging data in Alzheimer's disease: a systematic review**. Frontiers in aging neuroscience, 9, p.329. 2017.