LUNGS SEGMENTATION AND CLASSIFICATION USING MACHINE LEARNING IN CHEST X-RAY IMAGES

Ms. C. KEERTHANA Assistant Professor, Department of Computer Science Nallamuthu Gounder Mahalingam College, Pollachi, Tamilnadu.
Dr. B. AZHAGUSUNDARI Associate Professor, Department of Computer Science Nallamuthu Gounder Mahalingam College, Pollachi, Tamilnadu.

Abstract:
The identification and prediction of lung illnesses are increasingly vital and difficult. Image processing and machine learning techniques approaches are frequently used to detect early-stage of lung disease. Healthcare providers are still facing challenges in detecting cancer. The true origin of cancer and the complete treatment process remains a mystery.To identify the regions of the lungs affected by cancer, various image processing techniques are utilized. These techniques involve reducing noise, extracting features, detecting damaged regions, and comparing patient's medical history of lung cancer data. This research applies image processing and machine learning techniques to demonstrate accurate detection and prediction of lung cancer. The dataset of X-ray images used in the study was sourced from UCI repositories. To enhance the quality of the images, a local binary fitting mean filter was utilized during the image preprocessing stage. The K-means method was then applied to segment the images. This segmentation enabled the identification of the lung portion within the images. The data was classified using different machine learning algorithms, including RF, ANN, SVM and NB. The results revealed that the ANN model had the most accurate prediction of lung cancer. The main objective of the study is to segment the lungs portion and detect lung cancer with image processing and machine learning approaches.

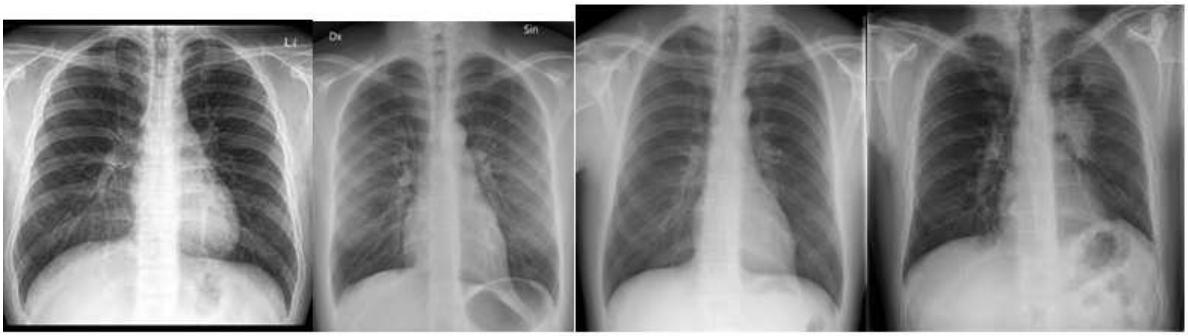
Keyword: Machine Learning, Lung Cancer, Median Filter, Segmentation, Chest x-ray images, Image processing

Introduction

Lung cancer, which is among the deadliest types of disease, causes approximately one million deaths every year. In the current state of medicine, it is essential to perform lung nodule detection on chest X-ray images. This is because lung nodules are becoming increasingly common and are the leading cause of this disease [1]. The images are transmitted into a computer, which then manipulates them to create a cross-sectional image of the body's internal organs and tissues [2].The American Cancer Society assesses that a patient's likelihood of surviving lung cancer is 47%, if it is discovered at an early stage. It is extremely improbable that lung cancer in its early stages will be accidently discovered on an X-ray image [6]. Lesions with a diameter of 510 millimeters or less that are spherical are notoriously difficult to find. Figure 1 displays an X-ray image of a normal and lung cancer.

In [3] developed a CAD approach with two experiments. In the first case, 300 X-ray images were used for computer simulation, while in the second, 888 X-ray images were used for patient-based ground truth. The method was developed using the LIDC-IDRI dataset, and images with wall thickness larger than 2.5mm were removed. The simulated pictures lung region was randomly implanted with spherical nodules of varied sizes. A 10-fold cross-validation method was used to analyze network hyper-parameterization and generalization in order to evaluate the CAD technique. The detection sensitivities were assessed in relation to the average number of false positives per image. By evaluating the mean and standard error between predicted and actual values, localization and diameter estimates accuracy was evaluated. Results regularly exhibited superior detection accuracy.

In many industries, image processing plays a vital role and it is used in X-ray scans of the lungs to identify malignant growths. Various image processing techniques such as noise reduction, feature extraction, identification of damaged areas, and comparison with the patient's medical history of lung cancer are employed to detect cancerous portions of the lung. Typically, digital image processing involves combining different aspects of an image to form a coherent object. To focus on a specific element of the overall lung imaging, this study employs a novel method. The split region can be observed in many different ways, including from various angles and when lighted in various ways. One of the major advantages of using this technique is being able to distinguish between areas of an image that have been affected by cancer and areas that have not been affected by cancer by contrasting the intensity of the two sets of images [6,7].



Normal Lung Images Cancer Lung Images

Figure 1:Chest X-rayimageoflungs.

Lung cancer is the most common cause of cancer-related death since the majority of patients acquire their diagnosis at a more advanced stage. Regardless of a country's level of industrialization or development, lung cancer is constantly regarded as one of the deadliest types of the disease. Effective control of lung disease relies on two factors: early detection of cancer and improving survival rates for those already diagnosed [8, 9].

A review of numerous methods for the classification and detection of lung cancer using image processing and classification may be found in the literature survey section. In the methods part, machine learning and image processing-enabled technology are used to accurately classify and predict lung cancer. Images are first acquired from the kaggle dataset. The Local Binary fitting mean filter is used to preprocess the image and as a result, image quality is enhanced. The K-means technique is then used to segment the images. The region of interest can be found using this segmentation. Then, methods for machine learning classification are used. The dataset and the outcomes of several strategies are described in the result section.

LiteratureSurvey

An active contouring model in Chest Radiology was proposed and implemented [1]. A variation level set function has been employed to segment the lungs. Proper segmentation of the lung parenchyma is essential for the diagnosis of lung disease. Significantly, the SBGF-New SPF function has been developed to segment X-ray lung images. This function is used to identify external lung limitations and to prevent errant expansion at the margins. Comparisons between the proposed algorithm and existing four different active contour models are being conducted. The results of these tests indicate that the strategy provided is robust and can be rapidly calculated [12].

Lung cancer was segmented by[13] using an active spline model for analysis. Through the application of this technology, segmentation of lung has been acquired using X-ray images. They employed a median filter to detect noise while preprocessing is being done. Additional K-means and fuzzy C-means clustering are employed for feature capture during the segmentation phase. The X-ray

image is segmented in this study before the final feature retrieval result is attained. The application of the SVM approach for classification led to the creation of the suggested model. MATLAB is used to simulate the results of the cancer detection system.

An IoT-based predictive model for predicting lung cancer using fuzzy cluster-based segmentation and classification was proposed by [11]. For accurate image segmentation, a fuzzy C-means clustering method was applied. The Otsu threshold approach was applied in this work to separate lung cancer images from transitional zones. To further enhance the segmentation representation, the right-edge picture is used with the morphological thinning technique. To accomplish incremental classification, we combine cutting-edge association rule mining (ARM), conventional decision trees (DT), and CNN with new incremental classification algorithms. The procedure was carried out using standard images from the database and current patient health data collected from IoT devices connected to the patient. The results of the research demonstrate that predictive model systems have improved in accuracy.

Developed a method for identifying the presence of lung cancer in an X-ray image using deep residual learning. The UNet and ResNet models were used to create a preprocessing pipeline by the researchers. The goal of this pipeline is to draw attention to and extract characteristics from malignant lung tissue. Predictions about the likelihood that aX-rayimage is malignant are gathered using an ensemble of XGBoost and random forest classifiers. The X-rayimage of malignant is calculated using the combined findings of the predictions made by each classifier. The LIDC- IRDI's accuracy is 84% greater than that of conventional methods [12]

In [14] developed an Optimal Deep Neural Network (OODN) to improve the classification of lung X-rayimage by reducing the number of features. Their approach proved more precise than existing techniques, and an automatic categorization method for lung cancer now saves time and eliminates human error. The researchers found that machine learning algorithms are now much better at distinguishing between normal and pathological lung scans. Based on the study's findings, the classification of lung images was successful, achieving 94.56% peer specificity, 96.2% accuracy, and 94.2% sensitivity. The study also showed that enhancing the efficiency of cancer detection in CAT scans is feasible.

The development of early detection and accurate diagnosis of lung cancer using CT, PET, and X-ray imaging by [9] in 2016 has attracted a great deal of attention and enthusiasm. The use of genetic algorithms, which enable early detection of lung cancer lesions by diagnosis, may lead to optimization of findings. Accurate and rapid classification of different stages in cancer images required the use of both naive Bayes and genetic algorithms. This was done to avoid complexity in the generation process. The accuracy of this classification is up to 80% [15].

In [19] used a preprocessing strategy to remove unwanted elements that are not affected by using median and wiener filters. Data quality is improved. The K-Means method is used for X-rayimage segmentation. EK-Mean clustering is a way to achieve clustering. Fuzzy EK average segmentation is used to extract contrast, homogeneity, area, correlation, and entropy features from images. We perform classification using a backpropagation neural network [20].

Neural ensemble-based detection(NED) is an automated method of disease diagnosis used in the study of [21]was proposed. Feature extraction, classification, and diagnosis were used as his three main components in the proposed approach. Chest X-ray films taken at Bayi Hospital were used in this experiment. This method is recommended due to its high needle biopsy detection rate and low false-negative identification. This automatically improves accuracy and saves lives [12].

In [13] developed a new early cancer detection algorithm that is more accurate than previous methods. This program uses technology to process images. Elapsed time is one of the factors considered when looking for anomalies in a photo of interest. The original photo shows the location of the tumor very clearly. To achieve better results, watershed segmentation and Gabor filter techniques are used in the preprocessing stage. The extracted zone of interest produces three phases (eccentricity, area, and perimeter) that are useful for detecting different stages of lung cancer. These phases can be found in the extracted zones of interest. Tumors are known to occur in a wide range of areas. The proposed method can provide accurate tumor size measurements at an early stage [15].

A computer aided lung classification method developed using artificial neural network was presented by Jinsa [6]. The parameters are calculated after the entire lung is segmented from the X-rayimage. The statistical parameters explained in this paper are used as features for classification. Different neural networks are used for the classification process. Thirteen training functions are employed for evaluating the performance of this system. The training function gives the highest accuracy rate.
In [7] proposed a lung cancer prediction system based on a deep learning technique called Google Net which shows better performance such as convergence rate, accuracy, sensitivity and specificity. Google Net is fine-tuned for the classification of lung cancer cells. This method used less training time and gave better results. Median intensity projection (MIPs) is also discussed in this paper which helps to learn features of cancerous and non-cancerous lung nodules that are compatible with the fine-tuned Google Net. This will increase the accuracy of the system when tested on the validation sets. After 300 epochs, accuracy of81%, sensitivity of 84% and specificity of 78% are produced by the trained system which is better than other available programs.

Methodology

This section demonstrates how machine learning and image processing-based technology may accurately classify and forecast lung cancer. Gathering dataset is the first step. After that, the lung chest images were preprocessed using a local binary fitting mean filter. Quality of the image is enhanced. The images are then segmented using the K-means technique. This segmentation makes it easier to identify the area of interest. Following that, machine learning-based categorization techniques are used. Figure 2 shows how machine learning and image processing technology are used to categorize and predict lung cancer.

The classification of images representing illnesses is greatly influenced by the image preprocessing. Images from chest X-ray can have a wide range of deviations including noise. Image filtering techniques can be used to eliminate these noises. To reduce the amount of noise, a local binary fitting mean filter is used in the image pre-processing stage [15].
This is achieved by reducing the size of the initial data matrix by the use of a technique called linear discriminant analysis (LDA). Examples of parallel transformation methods include the PCA and LDA. The PCA is an unsupervised analysis technique, as opposed to the supervised LDA technique. Latent dynamic analysis (LDA), in contrast to principal component analysis (PCA), looks for a feature subspace that increases the likelihood of class restoration. By giving the class restoration of the data a higher priority than the cost of processing, overfitting can be avoided [16].

The segmentation technique is applied during the processing of medical images. An image's primary function is to distinguish between elements that are helpful and those that are detrimental. In light of this, it divides a image into various parts according to how much each element resembles the elements around it. By adjusting both the texture and the intensity, this effect can be produced. A segmented area of interest can be used as a diagnostic tool to obtain information rapidly that is relevant to the current

problem. The method known as "K-means clustering" is the one that is most frequently employed when segmenting medical images.

The chest x-ray image is separated throughout the clustering process into a number of distinct groups, also known as clusters. There is absolutely no connection between these clusters. There are a few recognizable clusters seen in this image. Every single one of them has a distinct set of reference points to which each pixel is mapped. The K-means clustering algorithm divides the available information based on k reference points, dividing it into k distinct groups [17].

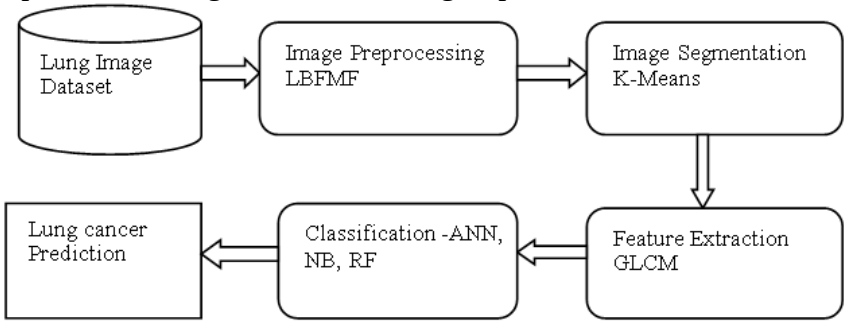


Figure2:Block Diagram of ClassificationandpredictionofLungDisease

In the medical field, artificial neural networks (ANN) are often used to sort medical images for disease detection. ANN works similar to the human brain when it comes to completing its tasks. By browsing a collection of already classified lung images, it is possible to obtain the necessary information to make an educated guess to which category the lung image belongs. This is possible by browsing the collection of classified images. All images in this image database are assigned to a category. Artificial neurons were designed to behave in the same way as their biological counterparts in the human brain make up an artificial neural network (ANN). Through connections, neurons can communicate with each other outside their bodies. Weights can be assigned to neurons and edges and can be changed at any time during the learning process.

There are three layers in an artificial neural network, namely the input layer, the hidden layer, and the output layer. This architecture is the most commonly used one. Though there are many different topologies that can be employed in artificial neural networks, the most popular ones include input, hidden, and final layers. There could be just one hidden layer, multiple hidden layers, or no hidden layers at all. All these possibilities are feasible. To achieve the desired output, the weights that need to be adjusted are hidden in a layer beneath the active layer [18]. The number of iterations and computing performance during ANN model training are closely linked. Having too few hidden layer neurons will decrease precision, while having too many neurons will increase training time.

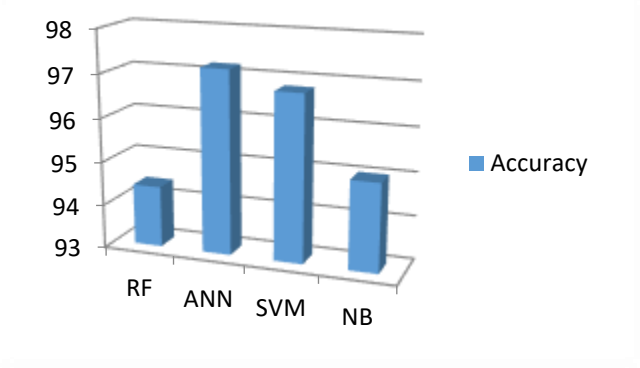


Figure 3:AccuracyofMachineLearningTechniques

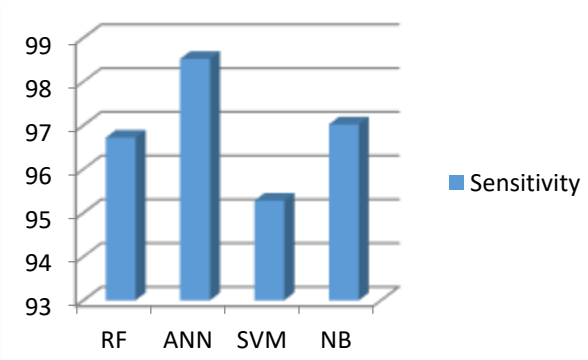


Figure 4: SensitivityofMachineLearningTechniques

The KNN technique is the most commonly used ML strategy for learning about ML algorithms. It is a type of supervised learning that doesn't use any parameters. The k-training NN phase is completed much faster than other classifiers, but testing takes longer and requires more memory. Before using k-nearest neighbors to classify new data points, pre-arranged data in various categories is necessary. In each labeled dataset, the algorithm can create a connection between x and y based on the provided training observations (x, y). At this point, it is typical to halt processing and determine the KNN function. In classification and regression models, neighbouring contributions may be prioritized, resulting in higher scores for those who live in close proximity to one another compared to those living farther away[19].

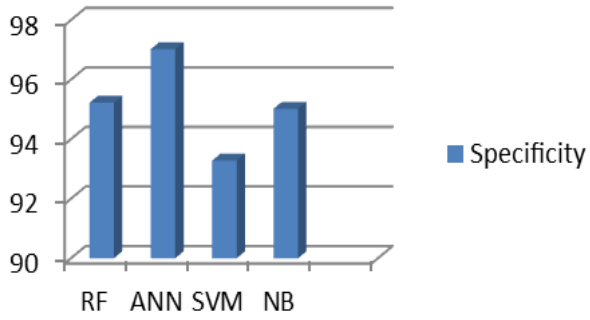


FigurE5:SpecificityofMachineLearningTechniques.

During testing, KNN had acceptable precision, but it was slower and more expensive in terms of memory and time. It requires a lot of memory to store the entire training dataset for prediction. Also, the features in the dataset with high magnitudes always hold a higher weight than those with low magnitudes due to Euclidean distance's sensitivity to orders of magnitude. Finally, KNN is not suitable for datasets with high dimensions. Many people use the random forest approach to build predictive models. RF can be used to achieve various applications, including regression and classification. [17].
By modifying datasets, it is feasible to develop machine learning algorithms that can accurately predict outcomes [18]. This approach is more user-friendly compared to other algorithms and is highly favored by the general public. Using this technique, a group of decision trees can be created, each trained differently. The current collection of trees, which represents different multiple-choice answers, was constructed using this method. These were then integrated to produce even more accurate estimates [20].

ResultAnalysis

In our experimental study, chest X-ray image dataset were collected from the UCI archive and around 150 chest x-ray images was collected. The training set used 70% of images and 30% of images was used as testing data. To implementMATLAB 9.0 was used.To enhance the image quality, preprocessing the images using the Local Binary Fitting Median filter was employed. Theoutput of the image was passed as input to the K-means technique was used to segment the images, to identify the region of interest (ROI). The output of the segmentation stage was passed as input to the machine learning classification methods to analyze the images.

For comparison of performance, the parameters used are accuracy,sensitivity,andspecificity.

True positives (TP): These are cases in which we predicted yes (they have the disease), and they do have the disease.
True negatives (TN): We predicted no, and they don't have the disease.
False positives (FP): We predicted yes, but they don't actually have the disease.
False negatives (FN): We predicted no, but they actually do have the disease.

The formula for the calculation of the accuracy is the addition of true positive and true negative (TP+TN) divided with respectively addition of true positive, true negative, false positive, false negative (TP+TN+FP+FN).

TP+TN-------- (1.1)

TP+TN+FP+FN-------- (1.2)

$Accuracy=equ1/equ2$ -------- (1.3)

$Accuracy$= (TP+TN)/(TP+TN+FP+FN) ------ (1.4)

$Sensitivity$= ($truepositives$ / $allactualpositives$) = (TP/ $TP + FN$) -------- (1.5)

$Specificity$= ($truepositives$ / $predictedpositives$) = TP/ $TP + FP$. --------- (1.6)

ResultsofdifferentmachinelearningpredictorsareshowninFigures3, 4 and 5. TheaccuracyofANNisbetter.

Conclusion

Lung cancer is a highly fatal disease that claims the lives of approximately one million people each year. Early detection of lung cancer is critical, and this can be achieved through the adoption of chest X-rays. Image processing is a crucial activity in various economic sectors. It is used to identify any malignant growths that may have developed in the lungs during X-ray imaging.To accurately predict and classify lung cancer, various techniques such as noise reduction, segmentation, classification and identification of damaged regions were employed. medical dataset is collected from UCI repository. This process involves the use of machine learning and image processing techniques. The first step is to collect images which are then preprocessed with a Local Binary Pattern Median filter to improve image quality. Finally, the K-means method is used to segment the lung portion from the image.To improve the accuracy of systems that detect lung cancer, this study uses algorithms for classification based on machine learning. The study found that artificial neural networks (ANN) were more effective in predicting lung cancer. By using robust categorization and prediction methods, the precision of these systems can be improved. In addition, this study provides modern images based on machine learning approaches that can be used for implementation needs.

Acknowledgement

The author acknowledges that the receipt of funding seed money from the management of Nallamuthu Gounder Mahalingam College, Pollachi for this research work.

References

[1] N. Deepa, B. Prabadevi, P. K. Maddikuntaet al., "An AI-basedintelligentsystemforhealthcareanalysisusingRidge-AdalineStochasticGradientDescentClassifier,"TheJournalofSuper-computing,vol. 77,no.2,pp.1998–2017, 2021.

[2] S.Chaudhury,A.N.Krishna,S.Guptaetal.,"Effectiveimageprocessing and segmentation-based machine learning techniques for diagnosis of breast cancer," Computational andMathematical Methods in Medicine, vol. 2022, Article ID6841334, 6pages,2022.

[3] A.HalderandA.Kumar,"ActivelearningusingFuzzy-RoughNearest Neighbor classifier for cancer prediction from micro-array gene expression data," Journal of Biomedical Informatics,vol.34, no.1,p.2057001, 2020.

[4] A. S. Zamani, L. Anand, K. P. Rane et al., "Performance ofmachine learning and image processing in plant leaf diseasedetection," Journal of Food Quality, vol. 2022, Article ID1598796,7pages,2022.

[5] S. Sandhiya and Y. Kalpana, "An artificial neural networks(ANN)basedlungnoduleidentificationandverificationmod-ule,"Medico-LegalUpdate,vol.19,no.1,p.193,2019.

[6] D. Palani and K. Venkatalakshmi, "An IoT based predictivemodelingforpredictinglungcancerusingfuzzyclusterbasedsegmentationandclassification,"JournalofMedicalSystems,vol.43,no.2,p.21,2019.
[7] S. Bhatia, Y. Sinha, and L. Goel, "Lung cancer detection: a deeplearning approach," in Soft Computing for Problem Solving,pp.699–705,Springer,Singapore, 2019.
[8] P. Joon, S. B. Bajaj, and A. Jatain, "Segmentation and detection of lung cancer using image processing and clustering techniques," in Progress in Advanced Computing and IntelligentEngineering,pp.13–23,Springer,Singapore,2019.
[9] E.E.NithilaandS.S.Kumar,"Segmentationoflungsfrom CT scanusingvariousactivecontourmodels,BiomedicalSignalProcessingandControl,vol.47,pp.57–62,2019.
[10] S.K. Lakshmanaprabu, S. N. Mohanty, K. Shankar,N.Arunkumar,andG.Ramirez,"Optimaldeeplearningmodelforclassificationoflungcanceron CTimages,"FutureGenerationComputerSystems,vol.92,pp.374–382,2019.
[11] J. Talukdar and P. Sarma, "A survey on lung cancer detection in CT scans images using image processing techniques," International Journal of Current Trends in Science and Technology,vol.8,no. 3,pp.20181–20186,2018.
[12] H.Z.Almarzouki,H.Alsulami,A.Rizwan,M.S.Basingab,H. Bukhari, and M. Shabaz, "An internet of medical things-basedmodelforreal-timemonitoringandavertingstrokesensors," Journal of Healthcare Engineering, vol. 2021, Article ID1233166, 9pages,2021.
[13] S.Chaudhury,N.Shelke,K.Sau,B.Prasanalakshmi,andM.Shabaz,"Anovelapproachtoclassifyingbreastcancerhistopathologybiopsyimagesusingbilateralknowledgedistillation and label smoothing regularization," Computational andMathematical Methods in Medicine, vol. 2021, Article ID4019358, 11pages,2021.
[14] T.Thakur,I.Batra,M.Luthraetal.,"Gene expression-assisted cancer prediction techniques,"Journal of Healthcare Engineer-ing,vol.2021, 9pages,2021.
[15] M. Heidari, S. Mirniaharikandehei, A. Z. Khuzani, G. Danala, Y. Qiu, and B. Zheng, "Improving the performance of CNN to predict the likelihood of COVID-19 using chest X-ray images with preprocessing algorithms," Int J Med Inform, vol. 144, no.5 September, pp:104-284, 2020
[16] E. Hussain, M. Hasan, M. A. Rahman, I. Lee, T. Tamanna, and M. Z. Parvez, "CoroDet: A deep learning based classification for COVID-19 detection using chest X-ray images," Chaos Solitons Fractals, vol. 142, p. 110495, 2021.
[17] A. M. Ismael and A. Şengür, "Deep learning approaches for COVID-19 detection based on chest X-ray images" Expert Syst Appl, vol. 164, p. 114054, 2021.
[18] Khan, Abdullah Ayub, Asif Ali Laghari, and Shafique Ahmed Awan. "Machine learning in computer vision: a review." EAI Endorsed Transactions on Scalable Information Systems 8.32, 2021.
[19] Sager, Christoph, Christian Janiesch, and Patrick Zschech. "A survey of image labeling for computer vision applications." Journal of Business Analytics 4.2, pp:91-110,2021
[20] Bayoudh, Khaled, et al. "A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets." The Visual Computer, pp: 1-32, 2021.