

Gabor-VGG19 Model for Enhancing COVID-19 Detection through Multimodal Fusion using CT and CXR Images

Gautami Shingan^{1*}, Shamik Tiwari^{2†} and Priya Ranjan^{3†}

^{1*}School of Computer Science, University of Petroleum and Energy Studies, Street, Dehradun, 248007, Uttarakhand , India.

²School of Computer Science, University of Petroleum and Energy Studies, Street, Dehradun, 248007, Uttarakhand , India.

*Corresponding author(s). E-mail(s): gautami.shingan@ritindia.edu;
Contributing authors: shamik.tiwari@ddn.upes.ac.in;
priya.ranjan@ddn.upes.ac.in;

[†]These authors contributed equally to this work.

Abstract

Multimodal approaches have gained significant attention in medical research, combining information from different imaging modalities. This paper introduces a multimodal approach for the detection of COVID-19 using Computed Tomography (CT) and Chest X-ray (CXR) images. Employing a deep learning-based pre-trained VGG19 model, our focus is on evaluating the classification performance of the proposed modified VGG-19 in comparison to single-modal CT images, single-modal CXR images, a multimodal fusion of CT and CXR images, and a multimodal mixture of CT and CXR images. For experimentation, an imbalanced dataset is employed and investigate the impact of multimodality on such datasets. Three augmentation strategies—no augmentation, partial augmentation, and complete augmentation—are implemented to assess their effects on the algorithm’s performance and check robustness with the local dataset. The results indicate that the multimodality fusion-based approach with partial augmentation consistently outperforms both single-modalities and the mixture of multimodalities. This research contributes valuable insights into the potential of multimodal approaches in enhancing disease detection models, particularly in the context of imbalanced datasets associated with COVID-19 diagnosis.

Keywords: Multimodal Fusion, Medical Images, Pre-trained Deep Learning Model, Early Fusion

1 Introduction

In early December 2019, COVID-19 was a worldwide epidemic triggered by the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2). The epidemic's impacts on economic and social aspects are unprecedented [1] [2]. As of early February 2022, the pandemic persists as a significant global concern, marked by nearly 6 million fatalities and more than 375 million confirmed infections. Alarming is a continuous rise in the worldwide count of infected individuals, with no discernible signs of abatement. Addressing the challenges of the COVID-19 pandemic requires the expansion of robust early diagnosis and intervention, coupled with effective mitigation measures. Consequently, the formulation of early detection approaches relies on the accurate identification and observation of these viruses in the spatio-temporal evolution [3].

Medical imaging and diagnostic modalities are categorized into two primary approaches [4]: The Unimodal and Multimodal approaches. The Unimodal approach relies on a single type of information, such as Ultrasound, MRI, CT, CXR images, text data, or audio data [5]. While this approach is straightforward and cost-effective, its limitation lies in accessing a restricted set of information, potentially reducing diagnostic accuracy. Unimodal models, relying on a single modality, may lead to false positives or false negatives, posing clinical risks. Complex cases may challenge these approaches. In contrast, Multimodal approaches combine data from various sources, enhancing diagnostic accuracy. Integration of image data with audio, patient records, or different image modalities (e.g., CT+CXR, CXR+clinical text reports [14], CXR+cough[2]CT+Ultrasound, CT+CXR+Ultrasound [5]) improves disease detection, identifies abnormalities early, and informs effective treatment strategies.

In this study, we diagnose COVID-19 using binary classification across CT and CXR modalities, analyzing performance in both single and multimodal settings. This research contributes valuable insights into the potential of multimodal approaches in enhancing disease detection models, particularly in the context of imbalanced datasets associated with COVID-19 diagnosis.

The construction of the paper is summarized as follow: Section II discussed Mixed modalities and Fusion modalities, reviewing the work conducted by researchers and addressing existing limitations. Section III describes the dataset details and proposed system; Section IV highlights the impact of argumentation over the datasets using single and multimodality approaches. Finally, Section V concludes the observations.

2 Literature Survey

A survey is carried out over two types of approaches such as Mixed modalities and Fusion modalities. Mixed modalities refer to the utilization of multiple types of imaging or data modalities within a study or analysis. (e.g., Chest X-ray and Computed Tomography) or additional non-imaging data (e.g., clinical information, patient history). Mixed modalities allow researchers to consider a broader range of information to enhance the understanding of a phenomenon or improve the performance of a model. On the other hand, In Fusion modalities involve the integration or combination of information from different modalities to gather various features associated with a particular disease. Fusion is a process where features or data from distinct modalities

are merged to extract complementary information or improve overall performance. In the context of medical imaging, fusion modalities often aim to enhance diagnostic accuracy by combining the strengths of different imaging techniques. Fusion methods may include early, late, or intermediate fusion, depending on when data from distinct modalities are combined. The review of the study is explained in the next sections.

2.1 Mixed Modality

Researchers evaluated augmentation techniques on VGG19 and ResNet50 models using the Cohen dataset, specifically the COVID-CT dataset [1]. In case of CT images, the VGG19 model achieved an accuracy of 84.87 %, outperforming the ResNet50 model, while, in the context of CXR images, the VGG19 model exhibited a remarkable accuracy of 99.35%, surpassing the ResNet50 model with an accuracy of 96.7%. These findings underscore the potential of VGG19 in handling CT images and highlight the overall strong performance of both models on CXR images. NasNetMobile[6] achieved an accuracy of 82.94% on CT images and 93.94% on CXR images. A transfer learning-based CNN model [7] was applied to the Cohen Dataset, achieving notable accuracies of 84.67% for X-ray and 98.78% for CT images. The proposed CNN architecture is highlighted as particularly effective for CT image datasets.

The study [8] utilized a mixed dataset of 1000 images, comprising 23 CT images and the rest from CXR modalities. Employing the VGG19 model with fine-tuning, the paper achieved a remarkable accuracy of 99%. [9] used a balanced small-size dataset containing 168 images in each class with total of 672 images. The proposed model is based on a tailored Deep Neural Network (DNN). Accuracy achieved by single modalities for CT and CXR is 95.83% and 96.13% respectively while mixed dataset CT and CXR together 96.28%. [10] the study employs CNNs on the COVIDx, Ozturk, and Ali Narin datasets, comprising 627 X-ray and 600 CT images. Despite class imbalances, CNN achieves impressive accuracies of 99.2% on CXR and 98.33% on CT scan images, with a combined accuracy of 99.5%. This multimodal approach enhances performance by 2%, The study showcases the model’s potential in medical image analysis, emphasizing its success in handling diverse imaging modalities. Notably, no fusion techniques were applied in their study, emphasizing the effectiveness of standalone models. The main limitation across these studies is the varying degrees of class imbalances and the absence of fusion techniques in some, leaving room for future investigations into their potential contributions to further improve diagnostic accuracy.

2.2 Fusion Modalities

An ensemble model combining ResNet50 and VGG16 with late fusion [11] was employed on the SARS-CoV-2 CT scan dataset. Individually, VGG16 achieved an impressive accuracy with 98.93% in CXR, while ResNet50 achieved 98.00% in CT. The ensemble model achieved an overall accuracy of 99.8%, highlighting the effectiveness of multimodality. MIDCAN model [12], incorporating a Convolutional Block Attention Module (CBAM) and employing late fusion, reported an accuracy of 98.02% with a small variation of $\pm 1.35\%$. CBAM allowed the AI model to focus on the lesion region,

enhancing its ability to discern critical features. A CNN model for COVID-19 detection using CXR images sourced from Cohen and local hospitals [2], the model achieved an accuracy of 94.75%, and the utilization of multimodality was noted to improve overall performance. Further in [13], CT/CXR images and clinical notes are Integrated using pre-trained models (VGG16, ResNet50, InceptionResNetV2, Mo-bileNetV2) on the Cohen dataset yielded 97.8% accuracy. ResNet50 and Inception-ResNetV2 demonstrated comparable performance, showcasing the effectiveness of multimodal fusion for enhanced COVID-19 detection. In [4], COVID-RDNet demonstrated 93% accuracy in CT, 98% in CXR, and a notable improvement to 99% accuracy when combining CT and CXR, showcasing a 6% performance enhancement through multimodal fusion. In [14], CBP-MMFN and DHP-MMFN models fused chest X-rays and clinical text reports from KMC Hospital and Indiana University datasets. Unimodal datasets achieved 79.22% accuracy for images and 90.40% for text. Multimodal datasets, combining image and text, achieved a higher accuracy of 97.35%, highlighting the superiority of multimodal approaches. [5] proposed a multi-modal graph attention model embedded approach for COVID-19 detection, utilizing CT, CXR, ultrasound, and text data. The Multi-Modal COVID-19 Pneumonia Dataset achieved an accuracy of 98.10%. The clinical data and CXR images [15] were reintegrated using pre-trained CNN models on the MIMIC-IV dataset. With a dataset size of 289, late fusion achieved an accuracy of 88.39%, while intermediate fusion reached 90.34%. As the dataset expanded to 1156, late fusion improved to 89.74%, and intermediate fusion increased to 91.30%. The findings suggest that multimodal fusion enhances COVID-19 detection accuracy, with performance improvements observed as the dataset size increases. Researchers across various studies have explored the efficacy of multimodal fusion in COVID-19 detection. The above studies underscore the advantage of multimodal approaches, emphasizing the potential for improved accuracy and diagnostic performance in detecting COVID-19 when collecting different image modalities.

3 Proposed Method

This paper proposed an early fusion technique to extract features such as color features and texture features from CT images and stack over the CXR modality pass through the VGG19 pre-trained model with dropout layers and finally classify images into two categories such as COVID and Non-COVID classes. The detailed system such as dataset details, modified VGG19, and fusion approach are explained in this section.

3.1 Dataset Details

The COVID-19 X-ray and CT Scan Image dataset is employed for experimentation, accessible at the following link: <https://www.kaggle.com/khoongweihao/covid19-xray-dataset-train-test-sets>. It encompasses two classes, Non-COVID and COVID diseases, featuring a total of 17,099 images across both CXR and CT modalities. Specifically, the CXR modality comprises 5,500 non-COVID images and 4,044 COVID images, while the CT modality includes 2,628 non-COVID images and 5,427 COVID images.

3.2 Proposed System

The increasing reliance on medical imaging for the of COVID-19 diagnosis has encouraged the exploration of deep-learning models for automated image classification. In this study, experimentation is carried out over the VGG19 pre-trained model due to its proven effectiveness in feature extraction tasks. The dataset is enhanced through pre-processing and augmentation techniques such as Width Shift, Height Shift, and Shear Range to address challenges related to image quality, uniformity, and class imbalance. Before feeding images into the VGG19 model, preprocessing steps are applied to standardize and optimize the input data. Resizing the images to a consistent size, such as 224x224 pixels, ensures compatibility with the VGG19 architecture. Normalizing pixel values to a standardized range, $[-1, 1]$, is crucial for mitigating variations in pixel intensities and enhancing the convergence of the deep learning model. Imbalanced datasets, particularly in medical image classification tasks, pose a challenge for model training. To address this, we employ augmentation techniques to artificially rise the size of the dataset and balance the different classes. This is crucial for achieving optimal performance during the fusion operation.

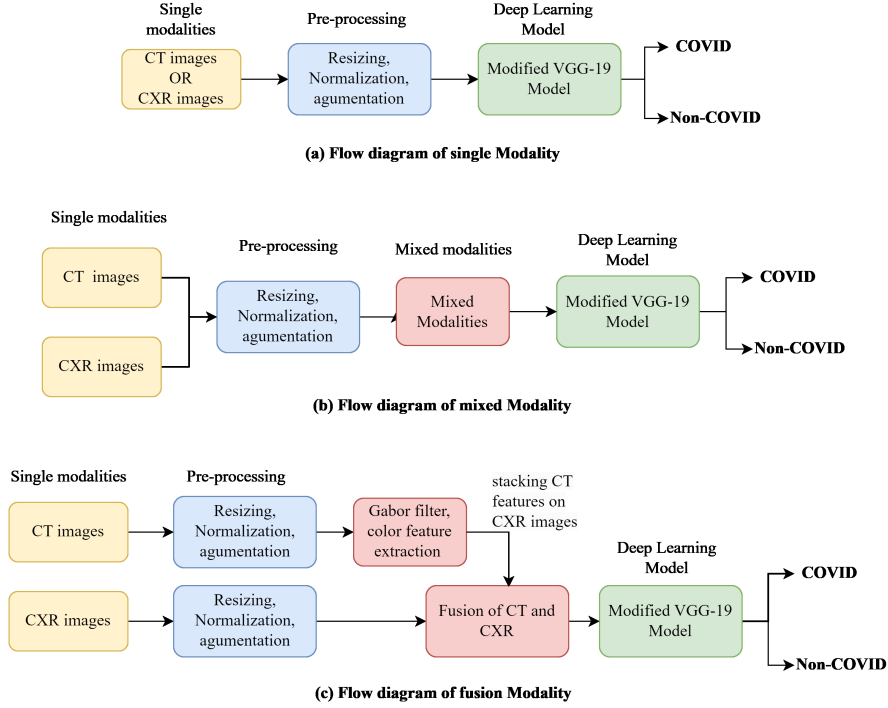


Fig. 1 The architecture of (a) Single modality (b)Mixed modality and (c) Fusion modality

The study explores disease classification using a VGG19 model in three modalities: single (Fig. 1a), mixed (Fig. 1b), and multimodal fusion (Fig. 1c). In the single modality, CT or CXR images undergo individual pre-processing and pass through

VGG19. The mixed modality combines independently pre-processed CT and CXR images before VGG19 classification. The multimodal fusion stacks features extracted from CT with CXR images for improved disease classification accuracy. The proposed model extends VGG19 with additional dense layers and dropout regularization (Fig. 2) to capture high-level representations and mitigate overfitting during training.

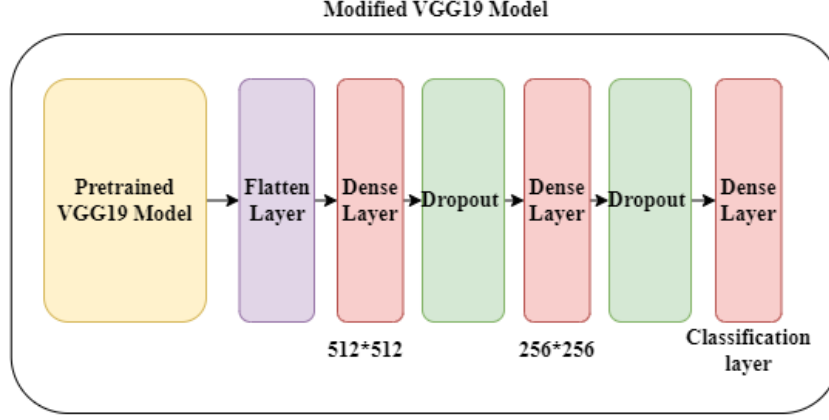


Fig. 2 Modified VGG19 Model

3.2.1 Image Fusion

Chest X-ray (CXR) and computed tomography (CT) represent two distinct imaging modalities, each possessing unique strengths and limitations in the context of medical imaging. CXR provides a quick overview but lacks detail, while CT offers high-resolution images but involves higher radiation. Fusing CXR and CT combines their strengths, providing a comprehensive understanding of chest pathology. The blend of broad structural context from CXR and fine-detail information from CT enhances diagnostic accuracy, which is particularly beneficial for complex conditions like COVID-19 synergistic approach facilitates more informed clinical decisions, showcasing the performance improvement achieved through multimodal integration. In the subsequent fusion steps, the CT image undergoes a grayscale conversion, simplifying its complexity while retaining essential information. The application of a Gabor filter extracts crucial texture features, revealing subtle patterns within the image. Concurrently, colour features are computed by averaging pixel values across different colour channels in the original CT image. The fusion is a collaborative process, where these distinctive features are extracted from CT images and harmoniously integrated with the information from chest X-ray images. (refer algorithm) This integration is achieved through stacking, resulting in a comprehensive representation that encapsulates both texture and colour information from both modalities. This fused image, enriched with a diverse set of features, is then subjected to normalization to ensure consistent pixel

values for subsequent analyses. Gabor filter is a linear filter used for detecting edges and texture patterns in images. The mathematical representation of the Gabor filter is:

$$G(x, y, \lambda, \theta, \psi, \sigma, \gamma) = \exp\left(\frac{x'^2 \gamma^2 y'^2}{2\sigma^2}\right) \cdot \exp\left(i\left(2\pi \frac{x'}{\lambda} + \psi\right)\right) \quad (1)$$

Where, $x' = x \cdot \cos(\theta) + y \cdot \sin(\theta)$ and $y' = -x \cdot \sin(\theta) + y \cdot \cos(\theta)$ are the coordinates of the point (x, y) over angle θ , λ is the wavelength of the sinusoidal factor, controlling the spacing of the parallel stripes of the Gabor function. ψ is the phase offset, σ is the standard deviation of the Gaussian envelope spatial aspect ratio, γ it controls the ellipticity of the filter. By varying the values of these parameters ($\gamma, \sigma, \theta, \lambda$ and ψ), different Gabor filters can be generated to capture various texture features in an image. The colour feature vector for the entire CT image can be calculated by taking the mean across colour channels at each pixel, which can be mathematically expressed as:

$$ColourCT(x, y) = \frac{1}{3}(CT(x, y)_R + CT(x, y)_G + CT(x, y)_B) \quad (2)$$

The colour features are calculated by taking the intensities of red, blue, and green in CT images over and coordinates.

4 Result Analysis

The section presents a thorough analysis of diverse modality-based classification approaches for disease diagnosis, encompassing single modality, mixed modality, and multimodality fusion strategies. The impact of different augmentation levels—no augmentation, partial augmentation, and complete augmentation—is explored across these frameworks using the VGG19 model as the core architecture. The experimentation followed defined parameters, including a batch size of 64, 50 training epochs, and standardized images to 224x224 pixels for compatibility with VGG19. The Adam optimizer with a learning rate of 0.0001 and binary cross entropy loss function were employed. Additionally, an augmented VGG19 architecture incorporated extra layers like Flatten, Dense with ReLU activation, and Dropout for regularization. The selection of Gabor filter parameters is crucial for enhancing the quality of the image and effectively learning exact patterns by minimizing noise and maximizing pattern discrimination. $\lambda = 5$ ensures that the filter is sensitive to medium-scale texture variations. This allows for the detection of important patterns without being overly sensitive to noise, thus improving the overall quality of the image. Set $\theta = [0, 4\pi, 2\pi, 4\pi]$ for covering a wide range of orientations, the filter becomes more robust in learning exact patterns regardless of their orientation. phase offset to ψ to 0 to extract both positive and negative features. Standard deviation δ is set to 2 which helps to capture fine details and avoid excessive smoothing. spatial aspect ratio of γ set to 0.5 to capture elongated texture patterns.

4.1 No Augmentation

To perform experimentation, the dataset was divided into training and testing sets with an 80-20 split ratio, allocating 80% of the data for training the model and reserving the remaining 20% for evaluating its performance shown in Table 1.

Table 1 Dataset distribution for No Augmentation

Classes	CT			Xray			Total
	Training	Testing	Total	Training	Testing	Total	
COVID	4341	1086	5427	3235	809	4044	9471
Non-COVID	2102	526	2628	4400	1100	5500	8128
Total	6443	1612	8055	7635	1909	9544	17599

Table 2 Performance Analysis over original Dataset

	Classes	Precision	Recall	F1-Score	Support	Accuracy
CT images over No augmentation						
Training Set	COVID	0.96	0.91	0.93	4341	90.81%
	Non-COVID	0.82	0.91	0.87	2102	
Testing Set	COVID	0.82	0.90	0.93	1086	90.38%
	Non-COVID	0.82	0.91	0.86	526	
CXR images over No augmentation						
Training Set	COVID	0.97	0.93	0.95	3235	92.82%
	Non-COVID	0.95	0.98	0.96	4400	
Testing Set	COVID	0.94	0.88	0.91	809	90.81%
	Non-COVID	0.92	0.96	0.94	1100	
Mixed Modality over No augmentation						
Training Set	COVID	0.95	0.97	0.96	4400	96.22%
	Non-COVID	0.97	0.95	0.96	4400	
Testing Set	COVID	0.93	0.97	0.59	1100	95.09%
	Non-COVID	0.97	0.93	0.95	1100	
Fusion Modality over No augmentation						
Training Set	COVID	0.97	0.95	0.95	4341	96.05%
	Non-COVID	0.95	0.97	0.96	4400	
Testing Set	COVID	0.96	0.93	0.94	1086	94.51%
	Non-COVID	0.94	0.96	0.95	1100	

The model is initially trained on the original imbalanced dataset, where CXR images outnumber CT images. Despite higher accuracy for CXR images (92.82%), testing accuracy remains comparable in both modalities. Surprisingly, the mixed modality outperforms the fusion modality when adopting multimodal approaches, potentially due to imbalanced class distribution. To address this, subsequent experimentation prioritizes partial augmentation techniques to enhance model performance on the imbalanced dataset.

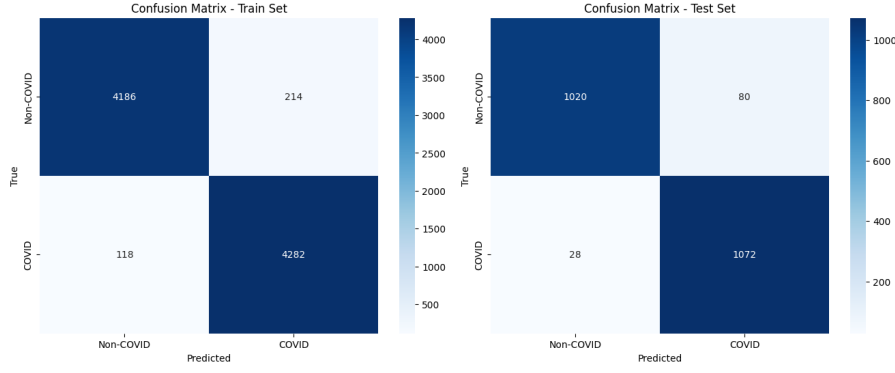


Fig. 3 Training and testing accuracy of Mixed modalities(No argumentation)

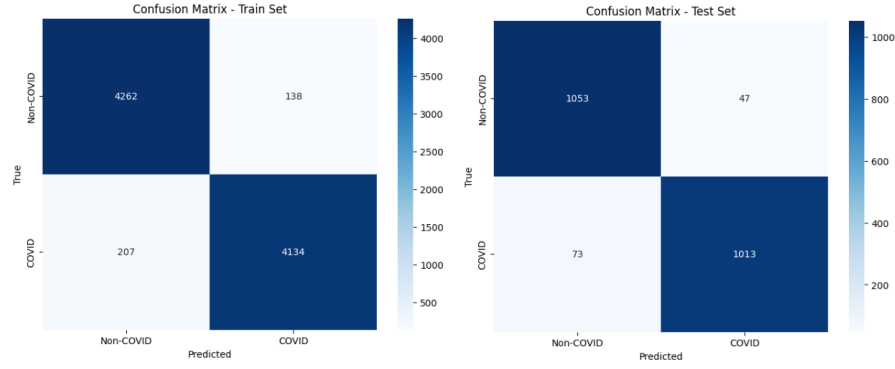


Fig. 4 Training and testing accuracy of Fusion modalities(No argumentation))

4.2 Partial Augmentation

The original dataset consists of 5427 COVID CT images, 4044 COVID CXR images, 2628 non-COVID CT images, and 5500 non-COVID CXR images. Augmentation involves adding 983 images to COVID CXR and 2872 images to non-COVID CT images. The same analysis is conducted on the partially augmented dataset.

Table 3 Dataset distribution for partial Augmentation

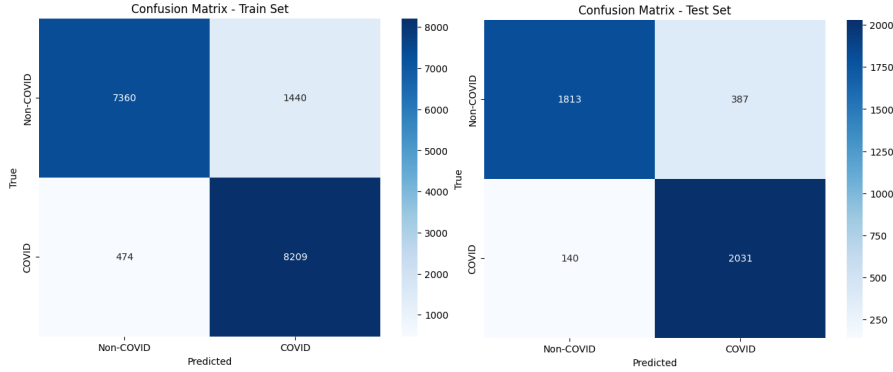
Classes	CT			Xray			Total
	Training	Testing	Total	Training	Testing	Total	
COVID	4341	1086	5427	4341	1086	5427	10854
Non-COVID	4400	1100	5500	4400	1100	5500	11000
Total	8742	2185	10927	8742	2185	10927	21854

Observations from Table V indicate that in a single modality, CXR produces higher accuracy than CT. In multimodality, fusion achieves the highest training accuracy of

Table 4 Performance Analysis over partial Augmented Dataset

	Classes	Precision	Recall	F1-Score	Support	Accuracy
CT images over Partial augmentation						
Training Set	COVID	0.85	0.96	0.90	4341	89.95%
	Non-COVID	0.96	0.84	0.89	4400	
Testing Set	COVID	0.84	0.96	0.90	1086	89.15%
	Non-COVID	0.96	0.82	0.88	1100	
CXR images over Partial augmentation						
Training Set	COVID	0.96	0.96	0.96	4313	96.09%
	Non-COVID	0.96	0.96	0.96	4400	
Testing Set	COVID	0.95	0.95	0.95	1086	95.01%
	Non-COVID	0.95	0.95	0.95	1100	
Mixed Modality over Partial augmentation						
Training Set	COVID	0.85	0.85	0.90	8683	89.05%
	Non-COVID	0.94	0.84	0.88	8800	
Testing Set	COVID	0.84	0.94	0.89	2171	87.94%
	Non-COVID	0.93	0.82	0.87	2200	
Fusion Modality over Partial augmentation						
Training Set	COVID	0.96	0.98	0.97	4341	96.92%
	Non-COVID	0.98	0.96	0.97	4400	
Testing Set	COVID	0.95	0.96	0.95	1086	95.37%
	Non-COVID	0.96	0.95	0.95	1100	

96.92% and testing accuracy of 95.37% on a balanced dataset. However, a negative impact is seen in the mixed dataset, where the absence of augmentation results in lower accuracy compared to the augmented dataset. This suggests that the lack of augmentation adversely affects performance with an increasing dataset size, while fusion modality enhances performance, especially on a balanced dataset.

**Fig. 5** Training and testing accuracy of Mixed modalities(Partial augmentation)

4.3 Complete Augmentation

A complete augmentation approach was conducted to validate the findings established in the partial augmentation, focusing on CXR Non-COVID images (5500 images).

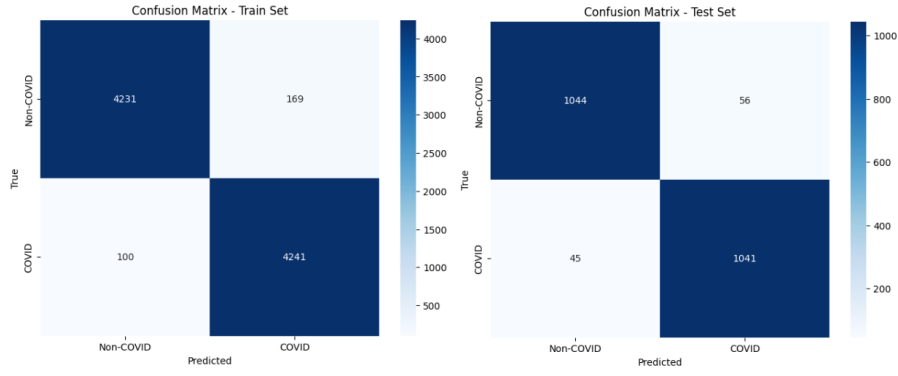


Fig. 6 Training and testing accuracy of Fusion modalities(Partial augmentation))

Uniform augmentation techniques were applied across all classes in both modalities, resulting in an augmented dataset with an increased image size of 5500.

Table 5 Dataset distribution for Complete Augmentation

Classes	CT			Xray			Total
	Training	Testing	Total	Training	Testing	Total	
COVID	4400	1100	5500	4400	1100	5500	11000
Non-COVID	4400	1100	5500	4400	1100	5500	11000
Total	8800	2200	11000	8800	2200	11000	22000

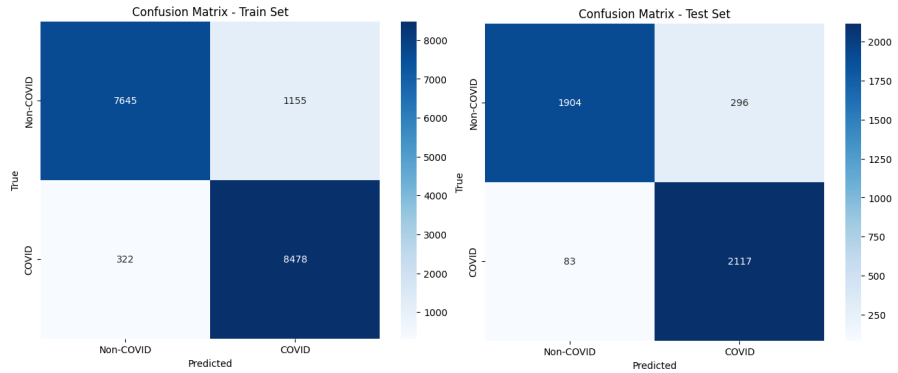
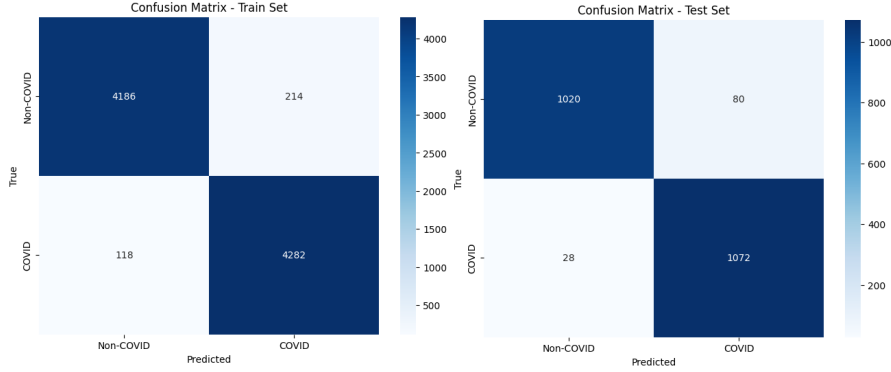


Fig. 7 Training and testing accuracy of Mixed modalities(Complete augmentation)

Based on the observations from Table 6, Single modality scenarios, CT outperforms CXR, especially in enhancing a balanced dataset. Transitioning to multimodality, fusion surpasses mixed modality, with a slight improvement in mixed modality compared to partial augmentation. This underscores the efficiency of integrating data from

Table 6 Performance Analysis over Completely Augmented Dataset

	Classes	Precision	Recall	F1-Score	Support	Accuracy
CT images over complete augmentation						
Training Set	COVID	0.91	0.94	0.92	4400	92.15%
	Non-COVID	0.94	0.90	0.92	4400	
Testing Set	COVID	0.92	0.91	0.92	1100	91.63%
	Non-COVID	0.91	0.92	0.92	1100	
CXR images over Complete augmentation						
Training Set	COVID	0.87	0.96	0.92	4400	91.15%
	Non-COVID	0.96	0.86	0.91	4400	
Testing Set	COVID	0.87	0.87	0.90	1100	90.54%
	Non-COVID	0.95	0.85	0.90	1100	
Mixed Modality over complete augmentation						
Training Set	COVID	0.88	0.96	0.92	8800	91.60%
	Non-COVID	0.96	0.87	0.91	8800	
Testing Set	COVID	0.88	0.96	0.92	2200	91.38%
	Non-COVID	0.96	0.87	0.91	2200	
Fusion Modality over complete augmentation						
Training Set	COVID	0.97	0.97	0.96	4400	96.22%
	Non-COVID	0.95	0.95	0.96	4400	
Testing Set	COVID	0.93	0.97	0.95	1100	95.09%
	Non-COVID	0.97	0.93	0.95	1100	

**Fig. 8** Training and testing accuracy of Fusion modalities(Complete augmentation))

modalities like CT and CXR, emphasizing the impact of augmentation techniques and balanced dataset creation on performance across different modality-based classification approaches. Further, the performance of the model is analyzed by plotting graphs of convergence in accuracy and loss over the training process. Fig. 9 illustrates the accuracy and loss trends observed during the evaluation of partial augmentation, showcasing its superior performance compared to complete augmentation. The graph suggests that fusion images, resulting from the merging of images before training and capturing crucial features swiftly during training, achieve optimal accuracy early in the process, outperforming single and mixed modalities. CT images exhibit rapid training due to their ability to highlight intricate features compared to CXR images.

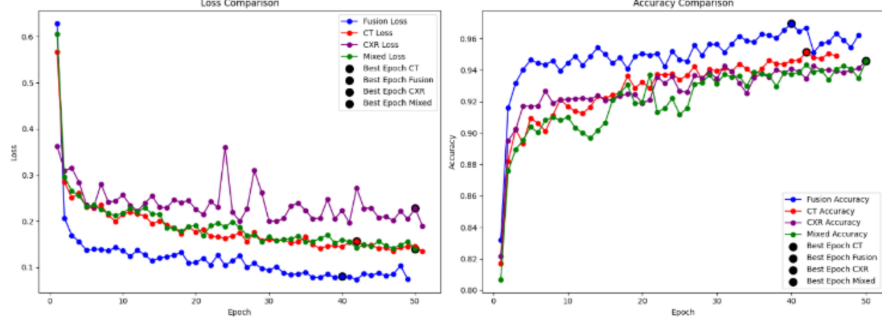


Fig. 9 Comparison of Accuracy and loss of single,mixed and fused modalities

Furthermore, when comparing mixed modality with fusion modality, it proves that fusion modality is significantly more efficient. The efficacy and generalization capability of the proposed Gabor-VGG19 model were evaluated using a local dataset. In this dataset, we have taken a multiclass dataset containing CT and CXR images of COVID, Pneumonia, Tuberculosis, and Normal images without performing augmentation and performed 2-class, 3-class, and 4-class classification, and the accuracies are discussed in Table 7.

Table 7 Dataset details of Multi-class classification

	CT images	Modalities CXR images	Total
COVID	2299	2616	4915
Normal	1728	2394	4122
Pneumonia	2000	1345	3345
Tuberculosis	1130	650	1780
Total	7157	7005	14162

Table 8 Performance Analysis over not augmented and partial augmented dataset

	2-class	Classes 3-Class	4-class
CT Modality	99.98%	97.67%	94.72%
CXR Modality	96.21%	96.08%	95.93%
Mixed CT+CXR Modalities	98.51%	96.26%	95.94%
Fusion without augmentation	96.71%	97.15%	96.30%
Fusion with augmentation	98.00%	98.00%	97.71%

Note: This table shows results of validation of proposed model over multi-class classification

CT modality-based images are fewer compared to those from the CXR modality. Consequently, CT scans demonstrate higher accuracy for binary classification but

perform less well for three-class and four-class classifications, whereas CXR images maintain approximately equal accuracy across all classification levels. When combining modalities, the number of classes impacts accuracy, with an increase in the number of classes leading to a decrease in accuracy. Fusion with augmentation is consistently beneficial because it ensures an equal number of CT and CXR images are used, which is crucial for efficient execution. The data also show that fusion is more effective with larger datasets. It has been confirmed that the proposed fusion approach outperforms both single and mixed modalities

5 Conclusion

The study explores single modality (CT or CXR), mixed modality, and multimodality fusion using a pre-trained VGG19 model, revealing insights into the impact of augmentation techniques and dataset balancing on classification performance. The study revealed that in the absence of augmentation, the CXR modality outperforms the CT modality in the single modality scenario, emphasizing the importance of pre-processing strategies in enhancing model accuracy. Furthermore, on a balanced dataset, the multimodality fusion approach consistently outperformed both single modalities and mixed modalities across all augmentation levels. Partial augmentation demonstrated performance improvements over the imbalanced dataset, highlighting the effectiveness of this strategy in addressing dataset imbalances. Complete augmentation further supported the superior performance of the CT modality in the single modality context and emphasized the continued success of the multimodality fusion approach. Finally, the model is robust against multiclass datasets. Overall, our findings underscore the efficiency of integrating data from various modalities and employing augmentation techniques, particularly in the multimodal fusion scenario, for enhanced disease detection models in the context of imbalanced datasets associated with COVID-19 diagnosis. A drawback of the current dataset is that CT and CXR images are not available for the same patient. In future, Performing image fusion using CT and CXR images from the same patient could be beneficial for the accurate diagnosis of lung diseases.

Data availability. The dataset is sourced from a Kaggle repository: <https://www.kaggle.com/khoongweihao/covid19-xray-dataset-train-test-sets>

Acknowledgements. I would like to express my gratitude to Rajarambapu Institute of Technology, Sakharale, for providing the necessary resources and support to implement this project.

References

- [1] Zouch, W., Sagga, D., Ehtioui, A., Khemakhem, R., Ghorbel, M., Mhiri, C., Hamida, A.B.: Detection of covid-19 from ct and chest x-ray images using deep learning models. *Annals of Biomedical Engineering* **50**(7), 825–835 (2022)
- [2] Kumar, S., Chaube, M.K., Alsamhi, S.H., Gupta, S.K., Guizani, M., Gravina, R., Fortino, G.: A novel multimodal fusion framework for early diagnosis and accurate

- classification of covid-19 patients using x-ray images and speech signal processing techniques. *Computer methods and programs in biomedicine* **226**, 107109 (2022)
- [3] Hu, F., Huang, M., Sun, J., Zhang, X., Liu, J.: An analysis model of diagnosis and treatment for covid-19 pandemic based on medical information fusion. *Information Fusion* **73**, 11–21 (2021)
 - [4] Fang, L., Wang, X.: Covid-rdnet: A novel coronavirus pneumonia classification model using the mixed dataset by ct and x-rays images. *biocybernetics and biomedical engineering* **42**(3), 977–994 (2022)
 - [5] Zheng, W., Yan, L., Gou, C., Zhang, Z.-C., Zhang, J.J., Hu, M., Wang, F.-Y.: Pay attention to doctor–patient dialogues: Multi-modal knowledge graph attention image-text embedding for covid-19 diagnosis. *Information Fusion* **75**, 168–185 (2021)
 - [6] Ahsan, M.M., Gupta, K.D., Islam, M.M., Sen, S., Rahman, M.L., Shakhawat Hos-sain, M.: Covid-19 symptoms detection based on nasnetmobile with explainable ai using various imaging modalities. *Machine Learning and Knowledge Extraction* **2**(4), 490–504 (2020)
 - [7] Chung, S.T., Morris, R.L.: A deep learning framework to detect covid-19 disease via chest x-ray and ct scan images. *International Journal of Electrical & Computer Engineering* (2088-8708) **11**(1) (2021)
 - [8] Kamil, M.Y.: A deep learning framework to detect covid-19 disease via chest x-ray and ct scan images. *International Journal of Electrical & Computer Engineering* (2088-8708) **11**(1) (2021)
 - [9] Mukherjee, H., Ghosh, S., Dhar, A., Obaidullah, S.M., Santosh, K., Roy, K.: Deep neural network to detect covid-19: one architecture for both ct scans and chest x-rays. *Applied Intelligence* **51**, 2777–2789 (2021)
 - [10] Goyal, L., Dhull, A., Singh, A., Kukreja, S., Singh, K.K.: Vgg-covidnet: A novel model for covid detection from x-ray and ct scan images. *Procedia computer science* **218**, 1926–1935 (2023)
 - [11] Hilmizen, N., Bustamam, A., Sarwinda, D.: The multimodal deep learning for diagnosing covid-19 pneumonia from chest ct-scan and x-ray images. In: 2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), pp. 26–31 (2020). IEEE
 - [12] Zhang, Y.-D., Zhang, Z., Zhang, X., Wang, S.-H.: Midcan: A multiple input deep convolutional attention network for covid-19 diagnosis based on chest ct and chest x-ray. *Pattern recognition letters* **150**, 8–16 (2021)
 - [13] Nasir, N., Kansal, A., Barneih, F., Al-Shaltone, O., Bonny, T., Al-Shabi, M.,

- Al Shammaa, A.: Multi-modal image classification of covid-19 cases using computed tomography and x-rays scans. *Intelligent Systems with Applications* **17**, 200160 (2023)
- [14] Shetty, S., Mahale, A.: Multimodal medical tensor fusion network-based dl framework for abnormality prediction from the radiology cxrs and clinical text reports. *Multimedia Tools and Applications* **82**(28), 44431–44478 (2023)
- [15] Kumar, S., Ivanova, O., Melyokhin, A., Tiwari, P.: Deep-learning-enabled multi-modal data fusion for lung disease classification. *Informatics in Medicine Unlocked* **42**, 101367 (2023)