International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Covid-19 and Viral Pneumonia Detection from Chest X-Ray Images Using Pretrained Deep Convolutional Neural Networks

Bambang Hartono Sinaga [1], Diaz D. Santika [2]

Submitted: 24/04/2023 **Revised**: 25/06/2023 **Accepted**: 08/07/2023

Abstract: The use of chest X-ray (CXR) in diagnosing respiratory diseases such as covid-19 and viral pneumonia has gained popularity due to its safe and non-invasive nature. However, interpreting CXR images as it is highly dependent on the expertise and experience of the radiologists may lead to inter-observer variability. A thorough analysis which in most cases takes some time then needs to be performed to get a final correct decision on whether a patient is indicated for a respiratory disease or not. Thus, research on devising more effective and efficient schemes to assist radiologists in the identification of respiratory diseases from CXR images is considered extremely important. In the previous studies, the convolutional neural network (CNN) based models demonstrate their capability to detect respiratory diseases not only they are found faster and reliable but also they are able to give a considerable high classification accuracy. In an attempt to obtain the general CNN based model capable of giving the best classification accuracy, sensitivity, and specificity on respiratory diseases from CXR images, in this paper powerful pretrained deep CNN models namely VGG16, DenseNet121, InceptionV3, Xception, and InceptionResnetV2 are computationally experimented on three different datasets. Dataset 1 is generated by combining images from Covid19 Radiography Database and Chest X-ray COVID-19 Pneumonia Dataset. Dataset 2 is created by combining images from Curated Chest X-ray Image Dataset, COVID-19 Pneumonia Normal Chest X-ray Images, and Chest X-ray COVID-19 Pneumonia Dataset. Meanwhile dataset 3 is taken from COVID-QU-Ex Dataset. State of the art accuracy of the pretrained CNN models is achieved by fine tuning the parameters of the convolutional layers of the base models and followed by feeding the high-level feature maps extracted from each corresponding base model into global average pooling (GAP) layer prior to classifying the respiratory diseases by fully connected layers. The highest average testing accuracy score as high as 99.2% is achieved by the InceptionV3 on multi-class CXR images.

Keywords*: Covid-19, CNN, Chest X-ray, Fine-tuning, Pneumonia.*

1. Introduction

The use of chest X-ray (CXR) in diagnosing respiratory diseases such as covid-19, tuberculosis, and viral pneumonia has gained popularity due to its safe and non-invasive nature. CXR provides a static image of the chest, allowing physicians and radiologists to assess lung morphology and detect pathological changes [1]. However, CXR interpretation can be challenging due to the complexity of the image features and the variability of visual interpretation. Interpreting CXR images is a subjective process that depends on the radiologist's expertise and experience. A study by Ali, et al [2] found that there is a considerable variation in the interpretation of CXR images by radiologists, and this may lead to diagnostic errors. Moreover, the use of CXR in detecting covid-19 especially in the early stage of the disease may result in low sensitivity and specificity [3]. To overcome these drawbacks, researchers have turned to convolutional neural networks (CNN) based models to develop automated methods for CXR interpretation. CNN based classifier models have shown promising results in identifying and diagnosing various respiratory diseases including viral pneumonia and COVID-19 with a considerable high classification accuracy and speed [4]. These models can learn complex image features and patterns from a large amount of data, reducing the subjective nature of CXR interpretation [5].

With the emergence of the convolutional neural networks (CNN) methodologies, the image features are automatically extracted and the learning process is carried out more deeply. CNN offer much better detection accuracy compared to the ordinary machine learning based approaches. Li, et al [6] presented CNN based model for diagnosing bacterial pneumonia, viral pneumonia, and covid-19. Employing datasets containg 1583 normal, 1493 viral pneumonia, and 305 covid-19 pneumonia cases, the model could achieve 89.1% classification accuracy. Another study by Jain, et al [7] employ deep learning model for covid-19 diagnosis using datasets of 6432 X-ray images including 1583 normal, 576 covid-19 and 4273 pneumonia cases. The model is capable of yielding the best accuracy, of 97.97%. Meanwhile, Sri, et al [8] use a CNN based method for detecting covid-19 and viral pneumonial from chest X-ray images. Employing datasets containing 3616 covid-19, 1345 viral pneumonia images, and 10192 normal cases, the model gives 91.39% detection accuracy. A similar

[1] *Computer Science Department, Binus Graduate Program, Master of Computer of Science, Bina Nusantara University, Jakarta 11480, Indonesia ORCID ID : 0009-0000-1264-4629*
[2] *Computer Science Department, Binus Graduate Program, Master of Computer of Science, Bina Nusantara University, Jakarta 11480, Indonesia ORCID ID : 0000-0002-9971-0744*
* *Corresponding Author Email: bambang.sinaga@binus.ac.id*

study by Farooq, et al [9] propose a deep learning model for covid-19 detection from chest X-ray images. They employ a dataset of 380 chest X-ray images consisting 180 Covid-19 and 200 normal (healthy). The model could obtained the best accuracy of 92.6%. The outcome of the aforementioned studies shows that the pretrained CNN models with their associated transfer learning are capable of achieving a considerable high classification accuracy of covid-19 and viral pneumonia detection from CXR images. In an attempt to obtain the general CNN based model capable of giving the best classification accuracy, sensitivity, and specificity on respiratory deseases from CXR images, in this study we performed an experimental computation on three different datasets using some powerful pretrained deep CNN models namely VGG16, DenseNet121, InceptionV3, Xception, and InceptionResnetV2. Dataset 1 is generated by combining images from Covid19 Radiography Database and Chest X-ray COVID-19 Pneumonia Dataset. Dataset 2 is created by combining images from Curated Chest X-ray Image Dataset, COVID-19 Pneumonia Normal Chest X-ray Images, and Chest X-ray COVID-19 Pneumonia Dataset. Meanwhile dataset 3 is taken from COVID-QU-Ex Dataset. State of the art accuracy of pretrained CNN models is achieved by fine tuning the parameters of the convolutional layers of the base models and followed by feeding the high-level feature maps extracted from each corresponding base model either into flatten layers or into global average pooling layers prior to classifying the class using fully connected layers.

2. Pretrained Deep Cnn Models

2.1. VGG16

VGG16 is a deep convolutional neural network (CNN) architecture, was introduced by Simonyan, et al [10] and has garnered significant attention in the field of computer vision for its exceptional performance in image classification tasks. With its 16 layers, including 13 convolutional layers and 3 fully connected layers, VGG16 has been widely employed and demonstrated remarkable accuracy and robustness in various domains, such as object detection, localization, and segmentation. State-of-the-art results on benchmark datasets like ImageNet have been achieved using VGG16, surpassing previous models. Moreover, VGG16 has been successfully utilized in medical imaging, where abnormalities in X-rays can be detected and aid in disease diagnosis. The model simplicity, as it utilizes small 3x3 filters, coupled with its ability to learn intricate image representations, contribute to its success. Additionally, VGG16 has been employed in transfer learning, where it has been fine-tuned for specific tasks, leading to improved performance and faster convergence on limited datasets.

2.2. DenseNet121

DenseNet121 is an influential deep convolutional neural network (CNN) architecture, was proposed by Huang, et al [11] and has attracted considerable attention in the field of computer vision. With its unique dense connectivity pattern, DenseNet121 addresses the vanishing gradient problem by establishing direct connections between layers, enabling enhanced information flow and feature reuse throughout the network. The architecture of DenseNet121 comprises 121 layers, including dense blocks and transition layers, which facilitate efficient feature extraction and dimensionality reduction. DenseNet121 has demonstrated outstanding performance in various computer vision tasks, including image classification, object detection, and semantic segmentation. It has achieved state-of-the-art results on benchmark datasets such as ImageNet, surpassing previous models in terms of accuracy and robustness. Moreover, DenseNet121 has been successfully employed in medical imaging applications, contributing to the detection and diagnosis of abnormalities in various modalities. The dense connectivity and transfer learning capabilities of DenseNet121 make it a versatile and powerful tool for advancing computer vision research and applications.

2.3. InceptionV3

InceptionV3 is a widely recognized deep convolutional neural network (CNN) architecture, was introduced by Szegedy, et al [12]. It has attracted considerable attention in the field of computer vision for its remarkable performance across various tasks. InceptionV3 employs a unique inception module, which allows for more efficient information processing and feature extraction through the use of multiple parallel convolutional operations at different scales. This architecture consists of 48 layers and demonstrates strong capabilities in image classification, object detection, and image segmentation tasks. InceptionV3 has achieved top results on benchmark datasets such as ImageNet, surpassing earlier models and showcasing its effectiveness in handling large-scale visual recognition challenges. Moreover, it has been successfully applied in domains like medical imaging, where it aids in the detection of diseases and abnormalities from various medical scans. InceptionV3 ability to capture intricate visual patterns, coupled with its versatility and performance, makes it a valuable tool for advancing computer vision research and applications.

2.4. Xception

Xception is an influential deep convolutional neural network (CNN) architecture, was introduced by Chollet [13] as an extension of the Inception architecture. It has garnered considerable attention in the field of computer vision for its exceptional performance across various tasks. Xception is characterized by its depthwise separable convolutions, which separate the spatial and channel-wise transformations, allowing for more efficient and expressive feature extraction. This architecture comprises a series of convolutional and depthwise separable convolutional layers, resulting in a highly efficient and parameter-efficient network. Xception has demonstrated remarkable performance in image classification, object detection, and semantic segmentation tasks, outperforming previous models on benchmark datasets such as ImageNet. Additionally, it has been successfully applied in various domains, including medical imaging, where it aids in the detection and analysis of diseases and abnormalities. The combination of depthwise separable convolutions, efficient architecture, and strong performance makes Xception a valuable tool for advancing computer vision research and applications.

2.5. InceptionResnetV2

InceptionResNetV2 is a deep convolutional neural network (CNN) architecture that combines the strengths of the Inception and ResNet architectures. It was introduced by Szegedy, et al [14]. InceptionResNetV2 builds upon the Inception architecture by incorporating residual connections, which enable the network to efficiently propagate gradients and alleviate the vanishing gradient problem. This architecture consists of a series of Inception modules and residual blocks, resulting in a highly expressive and powerful network. InceptionResNetV2 has demonstrated exceptional performance in various computer vision tasks, including image classification, object detection, and image segmentation. It has achieved state-of-the-art results on benchmark datasets like ImageNet, surpassing earlier models and showcasing its effectiveness in handling complex visual recognition challenges.Moreover, InceptionResNetV2 has been successfully applied in domains such as medical imaging, where it aids in the diagnosis of diseases and assists in medical research. The combination of the Inception architecture multi-scale feature extraction and the ResNet architecture skip connections makes InceptionResNetV2 a powerful tool for advancing computer vision research and applications.

3. Experimental and Result

3.1. CXR Image Dataset

In this study, three different publicly available CXR datasets were utilized to develop and evaluate the models. However, upon initial analysis, it was discovered that some of the classes in these datasets had a small number of samples, which could lead to a class imbalance problem. To overcome this challenge, these datasets were combined with other publicly available CXR datasets that contain similar classes of images. If the number of samples in a particular class is less than 3000 images, additional samples from other datasets with the same class were included. Dataset 1 is generated by combining images from the Covid-19 Radiography Database and the Chest X-ray covid-19 Pneumonia Dataset. The Covid-19 Radiography Database consists of 10192 normal, 1345 viral pneumonia, and 3616 covid-19 images. As the number of samples in the viral pneumonia class is less than 3000, additional samples were added to this class by including all the samples from the viral pneumonia class in the Chest X-ray Covid-19 Pneumonia Dataset. As a result, Dataset 1 contains 10192 Normal, 5618 Viral Pneumonia, 3616 covid-19. Dataset 2 is created by combining images from the Curated Chest X-ray Image Dataset, the Covid-19 Pneumonia Normal Chest X-ray Images, and the Chest X-ray Covid-19 Pneumonia Dataset. The Curated Chest X-ray Image Dataset has 3270 Normal X-Rays, 1656 viral-pneumonia, and 1281 covid-19 images. As the number of samples in the covid-19 and viral pneumonia classes is less than 3000, additional samples were added to the covid-19 class from all the covid-19 samples in the Covid-19 Pneumonia Normal Chest X-ray Images dataset. Similarly, additional samples were added to the viral pneumonia class from the viral pneumonia samples in the Chest X-ray Covid-19 Pneumonia Dataset. Consequently, Dataset 2 consists of 3270 Normal, 4657 Viral Pneumonia, and 3483 covid-19 images. Meanwhile dataset 3 is taken from COVID-QU-Ex Dataset, which contains 10701 normal, 11263 viral pneumonia, and 11956 covid-19 images. By combining these datasets, we were able to obtain a more balanced dataset for our models. This ensures that our models are not biased towards any particular class and can accurately classify all classes of CXR images. Each CXR dataset was divided into training, validation, and testing set as shown in Table 1.

Table 1. The dataset summary across different dataset, classes and train-validation-test splits

Dataset	**Set**	**Class**			**Total**
		Normal	**Viral Pneumonia**	**Covid-19**	
Dataset 1	Training	8,153	4,494	2,893	15,540
	Validation	1,020	562	361	1,943
	Testing	1,019	562	362	1,943
Dataset 2	Training	2,289	3,260	2,438	7,987
	Validation	490	699	523	1,712
	Testing	491	698	522	1,711
Dataset 3	Training	6,849	7,208	7,658	21,715
	Validation	1,712	1,802	1,903	5,417
	Testing	2,140	2,253	2,395	6,788

3.2. Data Preprocessing

Data preprocessing is an essential step to prepare the dataset before feeding it into model. Three different datasets have been used in this study, and various data preprocessing techniques have been applied to each dataset. For dataset 1, data augmentation has been used for the training data in the form of randomly transforming the input images. The transformations applied include rotation with a range of 15 degrees, rescaling the image to 1./255, shearing the image with a range of 0.1, zooming the image with a range of 0.2, flipping the image horizontally, and shifting the width and height of the image with a range of 0.1. The validation and testing data are only rescaled to 1./255. Additionally, the data are resized to a size of 224 x 224 with 3 channels. For dataset 2, the same data augmentation techniques have been used for the training data, but the zoom range is reduced to 0.1. The validation and testing data only undergo rescaling of 1./255, and the data are also resized to a size of 224 x 224 with 3 channels. For dataset 3, augmentation data has also been used for the training data, but the rotation range is reduced to 12 degrees, and only zooming and horizontal flipping are applied. The validation and testing data only undergo rescaling of 1./255, and the data are also resized to a size of 224 x 224 with 3 channels. By applying various techniques such as data augmentation and resizing, the dataset can be better fit to model, resulting in improved accuracy of predictions.

3.3. Model Training and Fine-tuning

In this study, the input data were preprocessed before being fed into five different pretrained deep CNN models, including VGG16, DenseNet121, InceptionV3, Xception, and InceptionResnetV2. Fine tuning was performed on a modified feature extractor and fully connected layer. In dataset 1 and 3, 50% of the layers were fine tuned, while in datasets 2, 25% of the layers were fine tuned. Each dataset also used a Global Average Pooling layer and fully connected layer. For the fully connected layer, 2 dense layers were added with 4096 units each and ReLU activation for dataset 1, and 2 dense layers were added with 1024 units each and ReLU activation for dataset 1 and dataset 3. For all three datasets, the output layer consisted of 3 units with softmax activation. The Adam optimizer with a learning rate of 1e-4 was used, and several techniques were employed, including early stopping, reducing learning rate, and model checkpointing. The models were trained for 20 epochs and used batch size of 32. Early stopping was set to monitor the loss, with patience of 5. Reduce learning rate was set to monitor validation loss, with patience of 3 and a factor of 0.5 for dataset 1 and 0.2 for dataset 2 and 3. Model checkpointing was used to save the best weights based on validation categorical accuracy. The models were fine-tuned to adapt them to the specific characteristics of the datasets, and the addition of fully connected layers aided in classifying the images into three categories. The early stopping technique was employed to prevent overfitting, while the learning rate reduction technique adjusted the learning rate based on the validation loss. Lastly, the model checkpoint technique was utilized to save the best model weights according to the validation accuracy.

3.4. Result And Analysis

3.4.1. Experimental Result on Dataset 1

Table 2 provides an overview of the performance metrics for different models, with each row representing a specific class. It is evident that all five models consistently achieved high accuracy values across all classes. Notably, the InceptionV3 model attained the highest accuracy of 0.992 (99.2%), while the DenseNet121model obtained the lowest accuracy of 0.983. Furthermore, when analyzing the precision values for each model,

it becomes apparent that they consistently exhibit high precision, ranging from 0.973 to 1.00. This suggests that the models effectively identify positive cases for each class while minimizing false positives. The recall values for each model also indicate strong performance, with values ranging from 0.942 to 0.998. These values imply that the models successfully identify a significant portion of positive cases for each class. Additionally, the F1-score values provide insights into the balance between precision and recall. Across the models, the F1-scores range from 0.966 to 0.994, demonstrating a favorable equilibrium between accurately identifying positive cases and minimizing false positives. Notably, the InceptionV3 model consistently achieved the highest accuracy and F1-score values across all classes, reinforcing its superior performance. On the other hand, the DenseNet121 model consistently yielded the lowest accuracy and F1-score values, indicating its comparatively weaker performance.

Table 2. Evaluation Results on Dataset 1

No	Model Name	Class	Precision	Recall	F1-Score	Accuracy
1	VGG16	Normal	0.982	0.994	0.988	0.987
		Viral Pneumonia	0.988	0.998	0.993	
		Covid-19	1	0.948	0.973	
2	DenseNet121	Normal	0.973	0.996	0.984	0.983
		Viral Pneumonia	0.996	0.986	0.991	
		Covid-19	0.991	0.942	0.966	

3	InceptionV3	Normal	0.989	0.997	0.993	0.992
		Viral Pneumonia	0.996	0.991	0.994	
		Covid-19	0.994	0.981	0.987	
4	Xception	Normal	0.988	0.997	0.993	0.991
		Viral Pneumonia	0.995	0.991	0.993	
		Covid-19	0.994	0.975	0.987	
5	Inception-ResnetV2	Normal	0.985	0.994	0.99	0.988
		Viral Pneumonia	0.991	0.986	0.988	
		Covid-19	0.992	0.975	0.987	

The ROC curves in Figure 1 provide the visual representation of the performance comparison among the models on testing data of dataset 1. The results indicate that all models exhibited excellent performance, with AUC values ranging from 0.9995 to 0.9999. Among the models evaluated, the InceptionV3 model attained the highest AUC value of 0.9999 (99.99%), this indicates that the InceptionV3 model achieves a better balance between true positive and false positive rates compared to the other models on dataset 1.

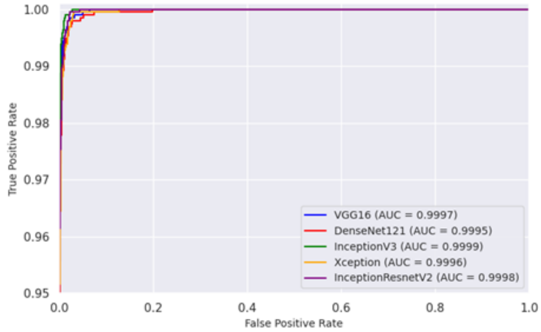

Fig 1. ROC Curve - Comparison of Model performances on dataset 1

3.4.2. Experimental Result on Dataset 2

Table 3 illustrated the evaluation metrics of five pre-trained models, namely Vgg16, Densenet121, InceptionV3, Xception, and InceptionResnetV2, on dataset 2. These models exhibit remarkable accuracy, ranging from 0.975 to 0.981, highlighting their proficiency in classifying image classes. In particular, InceptionResnetV2 achieves the highest accuracy at 0.981 (98.1%), showcasing its exceptional performance. Precision, which measures the ability to correctly classify instances within a specific class, is consistently high for all models across the three classes, ranging from 0.925 to 1. Furthermore, recall, which signifies the capacity to correctly identify class members, is impressively high for all models, ranging from 0.946 to 0.996. InceptionResnetV2 attains the highest recall for the normal and covid-19 classes, while Densenet121 achieves the highest recall for viral pneumonia. This further demonstrates the effectiveness of these models in accurately identifying instances from different classes. To evaluate the overall performance, the F1-score, a metric that balances precision and recall, is computed. The F1-scores for all models and classes are consistently high, ranging from 0.960 to 0.998. InceptionResnetV2 consistently achieved the highest F1-scores for normal and viral pneumonia class, while InceptionV3 achieved highest F1-score for covid-19 class. The results collectively establish the effectiveness of all five pre-trained models in accurately classifying normal, viral pneumonia, and covid-19 chest x-ray images. However, InceptionResnetV2 outperforms the other models, showcasing its superiority and suitability for this dataset.

Table 3. Evaluation Results on Dataset 2

No	Model Name	Class	Precision	Recall	F1-Score	Accuracy
1	VGG16	Normal	0.942	0.996	0.968	0.978
		Viral Pneumonia	0.994	0.953	0.973	
		Covid-19	0.994	0.996	0.995	
2	DenseNet121	Normal	0.925	0.998	0.96	0.975
		Viral Pneumonia	0.995	0.946	0.97	
		Covid-19	1	0.992	0.996	
3	InceptionV3	Normal	0.931	0.992	0.961	0.976
		Viral Pneumonia	0.988	0.95	0.971	
		Covid-19	1	0.996	0.998	
4	Xception	Normal	0.937	0.994	0.964	0.978
		Viral Pneumonia	0.994	0.953	0.973	
		Covid-19	0.998	0.996	0.997	
5	InceptionResnetV2	Normal	0.955	0.984	0.969	0.981

Viral Pneumonia	0.988	0.968	0.978
Covid-19	0.998	0.996	0.997

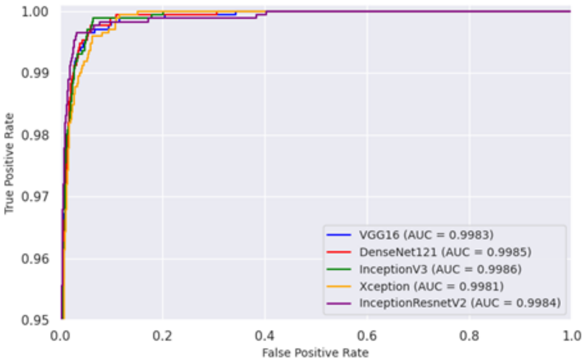

Fig 2. ROC Curve - Comparison of Model performances on dataset 2

The visual representation of performance comparison among the models on the testing data of dataset 2 can be observed through the ROC curves presented in Figure 2. The results clearly demonstrate that all models exhibited remarkable performance, as evidenced by their AUC values ranging from 0.9981 to 0.9996. Notably, the InceptionV3 model stood out among the evaluated models, achieving the highest AUC value of 0.9986 (99.86%). This outcome suggests that the InceptionV3 model outperforms the other models in terms of striking a better balance between true positive and false positive rates on dataset 2. Its superior AUC value indicates its ability to accurately classify instances from the testing data, ensuring a higher rate of correctly identifying positive cases while minimizing the occurrence of false positives. Thus, the InceptionV3 model exhibits strong discriminative power and robustness when applied to dataset 2, making it a highly suitable choice for the given task.

3.4.3. Experimental Result on Dataset 3

Table 4 presents an overview of performance metrics for different models, where each row corresponds to a specific class. It is evident that all five models consistently exhibited high accuracy values across all classes. Notably, the VGG16 model attained the highest accuracy of 0.958 (95.8%), while the Xception model achieved an accuracy of 0.949 (94.9%). Moreover, when examining precision values for each model, it becomes apparent that they consistently demonstrated high precision, ranging from 0.911 to 0.994. This suggests that the models effectively identified positive cases for each class while minimizing false positives. Xception model achieved the highest precision of 0.994 for normal class, DenseNet121 model obtained the highest precision of 0.948 for the viral pneumonia class, and VGG16 achieved the highest precision for covid-19 class.

Table 4. Evaluation Results on Dataset 3

No	Model Name	Class	Precision	Recall	F1-Score	Accuracy
1	VGG16	Normal	0.989	0.975	0.982	0.958
		Viral Pneumonia	0.942	0.953	0.947	
		Covid-19	0.941	0.945	0.943	
2	DenseNet121	Normal	0.989	0.967	0.978	0.953
		Viral Pneumonia	0.948	0.937	0.943	
		Covid-19	0.92	0.955	0.937	
3	InceptionV3	Normal	0.992	0.96	0.976	0.951
		Viral Pneumonia	0.945	0.938	0.942	
		Covid-19	0.914	0.954	0.934	
4	Xception	Normal	0.994	0.961	0.977	0.949
		Viral Pneumonia	0.94	0.938	0.939	
		Covid-19	0.911	0.948	0.929	
5	Inception-ResnetV2	Normal	0.986	0.969	0.977	0.953
		Viral Pneumonia	0.947	0.941	0.944	
		Covid-19	0.924	0.948	0.936	

The recall values for each model also indicated strong performance, with values ranging from 0.937 to 0.975. These values imply that the models successfully identified a significant portion of positive cases for each class. The VGG16 model achieved the highest recall of 0.975 for the normal class and 0.953 for viral pneumonia class, while the Densenet121 model obtained the highest recall of 0.955 for the covid-19 class. Furthermore, the F1-score values provided insights into the balance between precision and recall. Across the models, the F1-scores ranged from 0.929 to 0.982, demonstrating a favorable equilibrium between accurately identifying positive cases and minimizing false positives. The VGG16 model achieved the highest F1-score for all classes. These findings indicate that the VGG16 model consistently achieved the highest accuracy and F1-score values across all classes for this dataset, thereby reinforcing its superior performance. Figure 3 provides a visual representation of the performance comparison among the models on the testing data for dataset 3, as depicted by the ROC curves. The findings reveal that all models demonstrated impressive performance, with AUC values ranging from 0.9930 to 0.9951. Notably, the VGG16 model excelled among the evaluated models, achieving the highest AUC value of 0.9951 (99.51%). This outcome suggests that the VGG16 model surpasses the other models by achieving a more optimal balance between true positive and false positive rates on dataset 3. Its superior AUC value implies its capability to accurately classify instances from the testing data, thereby increasing the likelihood of correctly identifying positive cases while minimizing false positives. Hence, the VGG16 model exhibits robust discriminative power and reliability when applied to dataset 3, positioning it as an excellent choice for the given task.

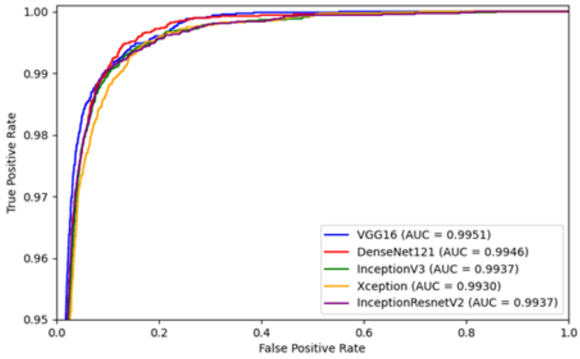

Fig 3. ROC Curve - Comparison of Model performances on dataset 3

3.5. Discussion

Table 5 presents a comprehensive comparison between our proposed models and other related works in the literature. The table contains important information such as the reference for each study, the datasets used along with the number of classes and sample size, the proposed methodology or model, and the reported accuracies. It should be noted that only related works with more than 1000 images were listed in Table 2, except for Oh, et al [15], who used four classes in total. Brima, et al [16] used the second largest dataset in Table 2, which consisted of a total of 21,165 X-ray images and achieved an impressive test accuracy of 94% using the VGG19 network architecture. On the other hand, Wang, et al [17] used the third biggest dataset, comprising of 13,962 X-ray samples, and achieved 93.3% test accuracy with their COVIDNet model. However, their dataset was not well-balanced, with covid-19 samples representing only 2.6% of the dataset. Similarly, Ozturk, et al [18] also used an imbalanced dataset and achieved 87% accuracy with their DarkCovidNet model. Narayan, et al [19], Apostolopoulos, et al [20], and Khan, et al [21] achieved higher classification accuracy with 98%, 98.75%, and 99.38% accuracy, respectively. However, it is worth noting that these studies evaluate their model performance using small samples of test images. Moreover, Khan, et al [21] reported the best accuracy of 99.38%, but their evaluation was based on only 200 test images belonging to two classes, 100 normal and 100 covid-19. In comparison, our proposed models were evaluated using three different datasets, each consisting of normal, viral pneumonia, and covid-19 X-ray images. In dataset 1, comprising 10,192 normal, 5,618 viral pneumonia, and 3,616 covid-19 X-ray images, we achieved an impressive accuracy of 99.20% using VGG16, DenseNet121, InceptionV3, Xception, and InceptionResnetV2 models. Similarly, in dataset 2, comprising 3,270 normal, 4,657 viral pneumonia, and 3,483 covid-19 X-ray images, we achieved an accuracy of 98.10% using the same models. Finally, in

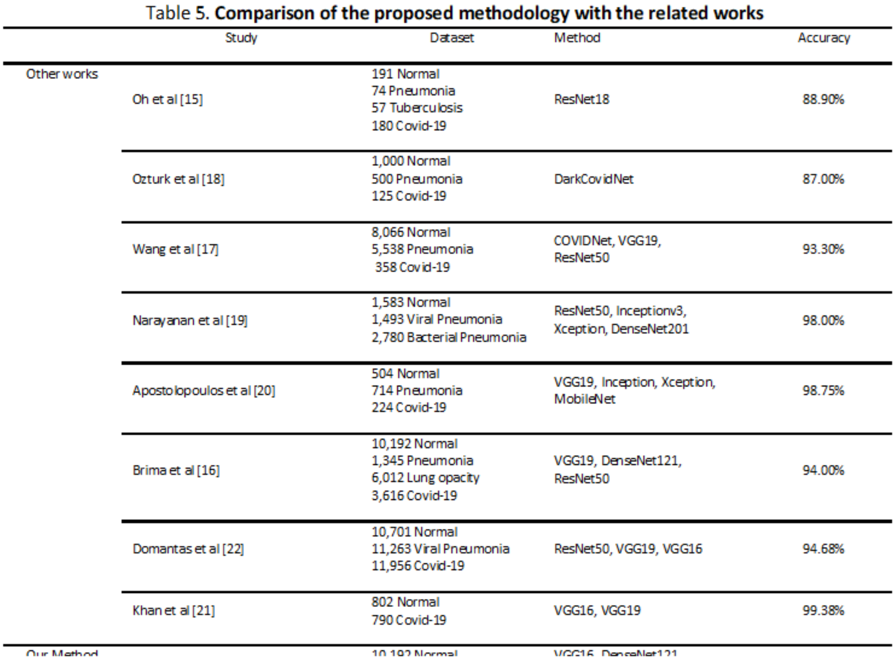

Table 5. **Comparison of the proposed methodology with the related works**

	Study	Dataset	Method	Accuracy
Other works	Oh et al [15]	191 Normal 74 Pneumonia 57 Tuberculosis 180 Covid-19	ResNet18	88.90%
	Ozturk et al [18]	1,000 Normal 500 Pneumonia 125 Covid-19	DarkCovidNet	87.00%
	Wang et al [17]	8,066 Normal 5,538 Pneumonia 358 Covid-19	COVIDNet, VGG19, ResNet50	93.30%
	Narayanan et al [19]	1,583 Normal 1,493 Viral Pneumonia 2,780 Bacterial Pneumonia	ResNet50, Inceptionv3, Xception, DenseNet201	98.00%
	Apostolopoulos et al [20]	504 Normal 714 Pneumonia 224 Covid-19	VGG19, Inception, Xception, MobileNet	98.75%
	Brima et al [16]	10,192 Normal 1,345 Pneumonia 6,012 Lung opacity 3,616 Covid-19	VGG19, DenseNet121, ResNet50	94.00%
	Domantas et al [22]	10,701 Normal 11,263 Viral Pneumonia 11,956 Covid-19	ResNet50, VGG19, VGG16	94.68%
	Khan et al [21]	802 Normal 790 Covid-19	VGG16, VGG19	99.38%
Our Method		10,192 Normal	VGG16, DenseNet121	

dataset 3, consisting of 10,701 normal, 11,263 viral pneumonia, and 11,956 covid-19 X-ray images, we achieved an accuracy of 95.80% using the same models. Additionally, it is worth mentioning the results from the study by Domantas, et al [22] which also evaluate the classification performance of ResNet50, VGG19, and VGG16 models on a dataset consisting of 10,701 normal, 11,263 viral pneumonia, and 11,956 COVID-19 X-ray images. They reported an accuracy of 94.68%, which is slightly lower than our results on the same dataset. It is important to highlight that our study utilized well-balanced datasets, where each class consisted of more than 3,000 images. This is in contrast to some of the related works in the literature that used imbalanced datasets, such as Wang, et al [17] and Ozturk, et al. [18] which might have affected their model performance. Our use of well-balanced datasets helped ensure that our models were trained and tested on a diverse range of images, leading to higher classification accuracy. These findings suggest that our proposed models are competitive and effective in the task of distinguishing between normal, viral pneumonia, and covid-19 chest x-rays images.

4. Conclusion

In conclusion, this study propose and evaluate pretrained deep Convolutional Neural Networks for the classification of chest x-ray images to aid in the diagnosis of covid-19, viral pneumonia, and other health conditions. The models were trained and fine-tuned using transfer learning on three different public datasets with well-balanced classes. The results demonstrated high accuracy, sensitivity, and specificity in distinguishing between covid-19 and viral pneumonia cases. The highest accuracy of 99.2% was achieved by InceptionV3 on dataset 1, followed by 98.1% by InceptionResNetV2 on dataset 2, and 95.8% by VGG16 on dataset 3. Comparison with related works in the literature revealed that the proposed models are competitive and effective. The utilization of transfer learning and fine-tuning techniques yielded favorable outcomes for the models. This study highlights the importance of using deep learning models in medical image analysis, particularly in the context of covid-19 and viral pneumonia diagnosis. The proposed model has the potential to assist medical professionals in making accurate and timely diagnoses, reducing the likelihood of errors and improving patient outcomes. For future research, there are several avenues to explore. Firstly, the impact of different data augmentation techniques should be investigated. Although the current study applies scaling, resizing, rotation, shearing, and zooming, it would be beneficial to analyze the effects of additional augmentation methods. This analysis could shed light on the specific techniques that yield the most significant improvements in Covid-19 and Viral Pneumonia diagnosis from CXR images. Additionally, further investigation into the fine-tuning process could be conducted. While the current study fine-tunes five pre-trained CNN models, namely VGG16, DenseNet121, InceptionV3, Xception, and InceptionResnetV2, on selective layers such as the feature extractor and output, it would be valuable to explore alternative pre-trained models and different selective layers for fine-tuning. This exploration could provide insights into the optimal combination of models and layers for achieving higher accuracy and efficiency in Covid-19 and Viral Pneumonia diagnosis. Moreover, the optimization process can be enhanced by exploring different optimization algorithms apart from the Adam optimizer. Comparing the performance of alternative optimization algorithms, such as RMSprop or Adagrad, may offer insights into potential improvements in the training process. Furthermore, the effectiveness of different callback methods could be examined. While the current study employs early stopping and learning rate reduction, investigating the impact of other callback techniques, such as batch normalization, may further enhance model performance and mitigate overfitting. Overall, further research should focus on investigating the effects of different data augmentation techniques, exploring alternative pre-trained models and selective layers for fine-tuning and evaluating different optimization algorithms and callback methods. These areas of study have the potential to enhance the proposed method and contribute to improved accuracy and effectiveness in diagnosing Covid-19 and Viral Pneumonia from CXR images.

Author contributions

Bambang Hartono Sinaga: Data curation, Training, Validation, and testing the models, Writing-Original draft preparation.

Diaz D. Santika: Data curation, Conceptualization, Methodology, Writing-Reviewing and Editing.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] R. K. Gupta, A. K. Yadav, and K. B. Gupta, "Medical imaging techniques and computer-aided diagnostic approaches for the detection of breast cancer using mammography: A review," J. X-ray Sci. Technol., vol. 27, no. 2, pp. 315-331, 2019.

[2] H. Ali, M. Abdar, M. Ali, and A. Awan, "Inter-observer variability in the interpretation of chest x-rays: A review," Cureus, vol. 12, no. 1, pp. e6633, 2020.

[3] H. Y. Wong et al., "Frequency and distribution of chest radiographic findings in patients positive for COVID-19," Radiology, vol. 296, no. 2, pp. E72-E78, 2020.

[4] E. E. Hemdan, M. A. Shouman, and M. E. Karar, "COVIDX-net: A framework of deep learning classifiers to diagnose COVID-19 in x-ray images," arXiv preprint arXiv:2003.11055, 2020.

[5] P. Rajpurkar et al., "Mura: Large dataset for abnormality detection in musculoskeletal radiographs," arXiv preprint arXiv:1712.06957, 2017.

[6] L. Li, L. Qin, Z. Xu, Y. Yin, X. Wang, B. Kong, J. Bai, Y. Lu, and Z. Fang, "Artificial intelligence distinguishes COVID-19 from community acquired pneumonia on chest CT," Radiology, vol. 296, no. 2, pp. E65-E71, 2020. [Online].Available:https://doi.org/10.1148/radiol.2020200905.

[7] R. Jain, M. Gupta, S. Taneja, and D. J. Hemanth, "Deep learning based detection and analysis of COVID-19 on chest X-ray images," Appl. Intell., vol. 51, pp. 1690-1700, 2020.

[8] N. Sri Kavya, T. Shilpa, N. Veeranjaneyulu, and D. Divya Priya, "Detecting COVID-19 and pneumonia from chest X-ray images using deep convolutional neural networks," Mater. Today Proc., 2022.

[9] M. Farooq, A. Hafeez, and M. Hassan, "A deep learning approach for COVID-19 diagnosis based on chest x-ray images," Expert Syst. Appl., vol. 164, p. 114054, 2021. [Online].Available:https://doi.org/10.1016/j.eswa.2020.114054

[10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[11] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), 2017, pp. 4700-4708.

[12] C. Szegedy et al., "Rethinking the inception architecture for computer vision," in Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), 2016, pp. 2818-2826.

[13] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), 2017, pp. 1251-1258.

[14] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-resnet-v2: Improved recognition for smaller networks," in Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), 2017, pp. 427-436.

[15] Y. Oh, S. Park, and J. C. Ye, "Deep learning COVID-19 features on CXR using limited training data sets," IEEE Trans. Med. Imaging, vol. 39, pp. 2688-2700, 2020.

[16] Y. Brima, M. Atemkeng, S. Tankio Djiokap, J. Ebiele, and F. Tchakounté, "Transfer Learning for the Detection and Diagnosis of Types of Pneumonia including Pneumonia Induced by COVID-19 from Chest X-ray Images," Diagnostics, vol. 11, p. 1480, 2021.

[17] J. Wang, Y. Peng, H. Xu, Z. Cui; Williams, R.O. The COVID-19 vaccine race: Challenges and opportunities in vaccine formulation. AAPS PharmSciTech 2020, 21, 1–12.

[18] T. Ozturk, M. Talo, E. A. Yildirim, U. B. Baloglu, O. Yildirim, and U. Rajendra Acharya, "Automated detection of COVID-19 cases using deep neural networks with X-ray images," Computers in Biology and Medicine, vol. 121, p. 103792, 2020. [Online]. Available: https://doi.org/10.1016/j.compbiomed.2020.103792.

[19] B. N. Narayanan, R. C. Hardie, V. Krishnaraja, C. Karam, and P. Davuluru VS, "Transfer-to-transfer learning approach for computer aided detection of COVID-19 in chest radiographs," AI, vol. 1, pp. 539-557, 2020.

[20] I.D. Apostolopoulos and T.A. Mpesiana, "Covid-19: Automatic detection from x-ray images utilizing transfer learning with convolutional neural networks," Phys. Eng. Sci. Med., vol. 43, pp. 635-640, 2020.

[21] I. U. Khan and N. Aslam, "A deep-learning-based framework for automated diagnosis of COVID-19 using X-ray images," Information, vol. 11, p. 419, 2020.

[22] K. Domantas and S. Clement, "The detection of COVID-19 in chest X-rays using ensemble CNN techniques," J. Med. Syst., vol. 46, no. 2, pp. 18, 2022.

[23] Prof. Parvaneh Basaligheh. (2017). Design and Implementation of High Speed Vedic Multiplier in SPARTAN 3 FPGA Device. International Journal of New Practices in Management and Engineering, 6(01), 14 - 19. Retrieved from http://ijnpme.org/index.php/IJNPME/article/view/51

[24] Ramesh, P. V. ., Hrishikesh, J. T. ., & Patil, M. S. . (2023). Infant’s MRI Brain Tissue Segmentation using Integrated CNN Feature Extractor and Random Forest. International Journal on Recent and Innovation Trends in Computing and Communication, 11(1s), 71–79. https://doi.org/10.17762/ijritcc.v11i1s.6002

[25] Jain, V., Beram, S. M., Talukdar, V., Patil, T., Dhabliya, D., & Gupta, A. (2022). Accuracy enhancement in machine learning during blockchain based transaction classification. Paper presented at the PDGC 2022 - 2022 7th International Conference on Parallel, Distributed and Grid Computing, 536-540. doi:10.1109/PDGC56933.2022.10053213 Retrieved from www.scopus.com