Chapter 3

Deep learning-based automated multi-class classification of chest X-rays into Covid-19, normal, bacterial pneumonia and viral pneumonia

3.1 Introduction

A global health crisis, Covid-19, began in December 2019, which was so impactful that it was declared a pandemic by the WHO [22]. It forced us to do severe lockdowns and strict rules for movements across the countries. It has changed a lot of things, including lifestyle, work culture, economic depression, and high mortality rate all over the world, which has led to a massive burden on health infrastructure, which collapsed due to the enormous number of patients and lesser medical resources available.

Novel coronavirus attacks the respiratory system of the human body. It infects the lungs, due to which the patient is unable to breathe correctly and eventually dies due to lack of oxygen to the lungs, which has several symptoms [23] such as fever, cough, and weakness. However, these systems vary from individual to individual based on their immune system response; along with that, loss of taste and smell is also another symptom. With an increasing infection rate of Covid-19, the patient suffers chest pain and loss of breathing. To assist that, patients need urgent medical oxygen supply to restore breathing.

As of now, there is no particular drug that can cure the disease; instead, we treat it using symptomatic treatment. However, we might not have any symptoms in several cases [24]. There are several vaccines [25] in the market; none proved to be having 100% efficacy in dealing with coronavirus. Even vaccinated people get affected with Covid- 19 [26] but with lesser severity and risk to lives. The main challenges to deal with it – rapid spreading continuously and mutation into several of its variants. Recent coronavirus variants are delta and omicron, which are deadly compared to earlier variants. Researchers have not yet tested vaccine effectiveness on omicron.

A better way to deal with Covid-19 is to make an early diagnosis as soon as possible, which can initiate treatment, and we can stop or localize the spread of Covid-19 into a few persons and regions. This virus mainly spreads through human-to-human interaction, and several researchers also found that it is transmitted through the air as media as well. Any surface or object can easily share this with another person. Being an invisible threat and hardly curable disease, it has affected almost all of the world population from all age groups, from small children to the oldest people. Nobody could avoid it, even after being quarantined. It shows how deadly it is for humankind, and hence bulk testing of Covid-19 will enable us to find Red-Zones where patients and spread rate and mortality rates are higher. We can mark them and implement a lockdown in that region and quarantine its people. We need a method to be reliable, fast and provide results within less time.

Out of several available methods [27] such as RT-PCR, INAT, antibody test, serology test, and medical imaging, Covid-19 testing uses the traditional diagnosis RT-PCR-based method, which eventually takes more time (roughly 4–6 hr). The higher spread rate, infection rate, and very high mortality rates lead to finding alternative solutions to the diagnosis at tiny fractions of time. To do this, several pathologists started taking the help of Chest X-Ray and CT imaging as primary diagnostic tools because these tools are widely available across the country around each corner. Also, considering the nature of the virus mutating over time, it has been the primary requirement to diagnose it as early as possible.

Along with the RT-PCR report, X-Ray and CT scans have also been helpful for primary diagnosis of virus and its severity based upon CT score [28]. With a CT-score, patients are diagnosed as either Normal or Covid-19 and live-in quarantine or in-home isolation, but in several cases, it has been found that Viral Pneumonia and other flues go undetected in the early stage. It gives False-Negative results, which is dangerous for the patient, as they do not have any symptoms and CT score is negative, yet they might be contaminated and can lose their lives. Hence, CT and X-Ray have certain limitations which need to be taken care of: a) It is unable to differentiate between Covid-19 and other respiratory infections. b) Most people might get a Normal CT Score when there is low severity, which is not accurately identifying a patient's condition as an early diagnosis. c) Due to its rapid spread based on touching objects and human-to-human interactions, healthy people and health care staff might also be affected while taking X-Ray of infected patients. Compared to CT, X-Ray is widely available, even in remote areas in most countries. Hence, we have utilized X-Rays for this study instead of CT However, CT is a better and advanced imaging modality than X-Rays in terms of image quality and angle of visualization.

We introduced a Deep Transfer learning-based [29] method that utilizes its ability to classify input X-Ray images into Covid-19, normal, viral Pneumonia, and Bacterial Pneumonia. To achieve this, we have applied the Inception module and the VGG-16 Net, compared against the pre- existing VGG-16 Net, Res-Net, and Inception-Net. For the proposed method, we achieved accuracies for each class in Case (01) – Covid-19 (91.86%), Normal (84.11%), Pneumonia Bacterial (83.91%), and Pneumonia Viral (94.77%). For Case (02) – Covid-19 (97.67%), Normal (97.93%), Pneumonia Bacterial (98.19%), and for Case (03) – Covid-19 (99.61%) and Normal (99.61%). As discussed in the results section, we found better performance figures than pre-existing methods.

3.2 Related Works

This section will review several methods and studies based on medical image classification, which give us insights about moving forward to conduct our research work.

3.2.1 Deep learning in medical imaging

With the latest improvements and advancements in computational capacity and the availability of massive datasets, researchers have developed various deep learning-based algorithms that outperform the specialists in that field, such as medical image disease detection, classification, and segmentation. Due to this, we have been able to move towards an automated diagnosis of several diseases, for example, brain tumor detection, skin lesion classification, breast cancer detection, and the severity proportion of any diseases like Covid-19. Several deep learning architectures, such as ALEX-Net [30], Le-Net [31], Google-Net [32], VGG-16 Net [2], U-Net [33], and so on, have been used based upon the type of disease or task. One single deep learning architecture is not sufficient for all, and hence it gives varying performances over different tasks. Also, not all the time do we need the same performance metrics to evaluate our model, so a thorough check is necessary while designing and implementing the model.

Developing a raw model is challenging and requires vast computational resources, which might not be available at all institutions. But there is a way to deal with it; we can utilize pre-trained models, trained over millions of images. We can use their weight parameters up to the second last layer of the model and tweak the final few layers based on our requirements.

3.2.2 Covid-19 Classification

It has been 2 years since the first reported case of Covid-19 in December 2019. Deep learning-based image classification has been in existence for so long, which has been utilized by many researchers to deal with Covid-19 classification with the help of medical images – Chest X-Ray and CT.

Wang et al. suggested one prevalent approach to classification [34]in the form of the proposed Covid-19-Net. They declared the open-source network design for Covid-19 classification using Chest X-Ray as the first of its kind. They also provided an open-source Covid-19X dataset.

Another notable work by Rajpurkar [35]. developed an algorithm that could beat the radiologists to detect Pneumonia from the Chest X-Ray images. They utilized a public dataset of chest X-Rays and detected 14 diseases better than radiologists. It was a revolutionary work as it surpassed the radiologist's performance.

A machine-learning-based approach was also applied by Kassani et al. (2020)[36], in which they tried several deep learning-based models and machine learning classifiers to classify the results. Arias-Garzón et al. (2021)[37] discussed Convolutional Neural Network (CNN)-based Covid-19 detection in X-Ray images. They studied pre-existing VGG-16 and U-Net to process the Chest X-Ray and classify them as positive or negative for Covid-19. This process utilized pre-processing to remove unwanted chest X-Ray portions from the image using the lung segmentation method. The best of their model achieved Covid-19 detection accuracy of around 97%.

Das et al. (2021)[38] presented ensemble learning with a CNN-based approach. Their work has adopted multiple state-of-the-art models such as DenseNet201, Res-Net50V2, and InceptionV3. They have used a smaller dataset of 538 Covid-19 positives and 468 Covid-19 negative cases. Their approach gave a classification accuracy of 91.62%. We improve these results with a significant increase in the dataset and by tuning training parameters.

Maghded et al. (2020)[39], in one of their works, tried to create a novel AI-enabled framework to Covid-19 diagnosis using smartphone. In another notable work by Maghdid et al. (2020)[40], the authors performed data collection via different online resources as well as aimed to build one Covid-19 detection algorithm. To do that, they have utilised simple CNN along with pretrained Alex-Net and achieved accuracy around 94.1%.

Shah et al. (2022)[41], in their work, gave insights about the Covid-19 and its research challenges, which gave a good review about various factors associated with Covid-19.

Maghdid et al. [42] in one of their work designed a smartphone-based approach to manage Covid-19 lockdown by finding the contamination zones .

Ferhat Ozgur Catak et al. [43] in one of their latest research articles utilised a transfer-learning-based CNN model for Covid-19 detection. They proposed an uncertainty quantification-enhanced transfer-learning-based CNN to predict the presence of Covid-19. They achieved approx. 75% accuracy.

Based on the above literature survey, we found that there is a requirement for multiclass classification instead of just binary classification. This paper is an attempt to perform multiclass classification in three separate case studies to make our model more versatile, and we can use this model as per our requirement and urgency of the situation. Our model performed best for binary class classification 99.95%. It also performed very well for other case studies which involve multiclass classification. For four-class classification, we achieved an accuracy of 87.32% and for three-class classification, we achieved 96.89%.

3.2.3 Research Gaps

After reviewing the literature, we have found out few common points that we need to take into consideration-

1. Data insufficiency to train an extensive deep neural network to accurately classify into various classes.
2. Most of the studies rely upon either just CT or CT with CXR
3. There is a need to develop a model that gives better classification results on a smaller dataset that addresses multi-class classification.

To deal with the problem of data scarcity, we have utilized the collected dataset from different resources and then cleaned it as per our requirements. To enhance the size of our dataset, we have performed some data pre-processing and data augmentation. We stuck to Chest X-Ray for input image type mainly because it is widely available in most regions and cost-effective. We utilized it for Covid-19 detection using deep learning as early as possible.

3.3 Dataset

3.3.1 Original collected Dataset

The Curated Covid-19 X-Ray Dataset [35] has been collected, assembled, and maintained from version 1 to version 3 by a joint effort between the Indian Institute of Science,PES University, MS Ramaiah Institute of Technology, and Concordia University. We have used version 3 of this dataset. This dataset collects Covid-19 X-Ray Dataset from 15 publicly available datasets cited in the given reference.

The collected dataset includes four classes, namely— 1. Covid-19 X-Rays, 2. Normal X-Rays, 3. Viral Pneumonia X-Rays, and 4. Bacterial Pneumonia X-Rays. As shown in the Figure.3.1, there are total of 4558 Covid-19 X-Rays, 5403 Normal X-Rays, 4497 Viral pneumonia X-Rays, and 5768 bacterial pneumonia X-Rays. Out of which, 1379 Covid-19 X-Rays, 1476 normal X-Rays, 2690 viral pneumonia X-Rays, and 2588 bacterial pneumonia X-Rays are duplicates based on the image similarities, thus are removed. Several other images that were defective or poor-quality images were also removed, and the final present dataset used in this study is as shown in Figure.3.2.

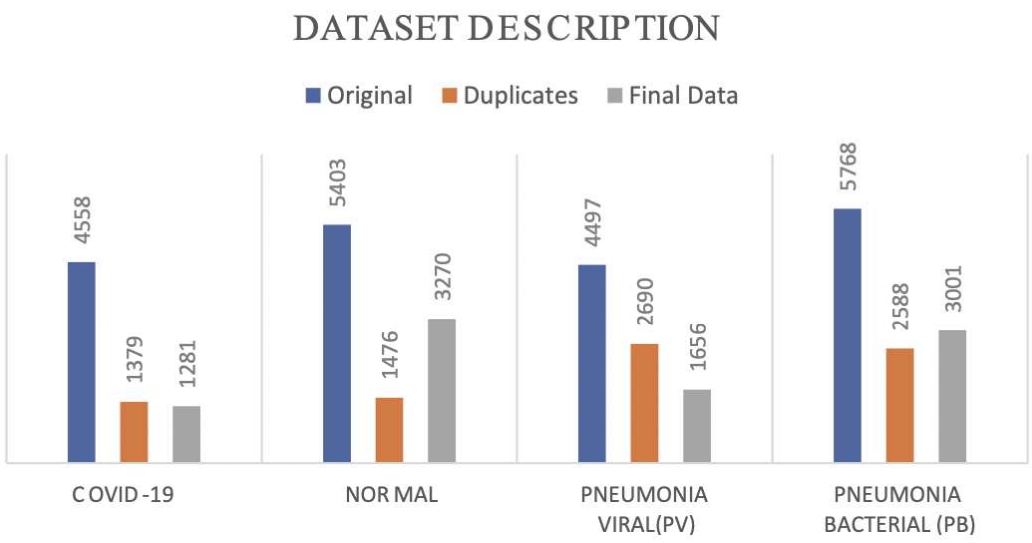

FIGURE 3.1: Original dataset

In the present study, we have studied the behavior of the same model to evaluate classification accuracy over-

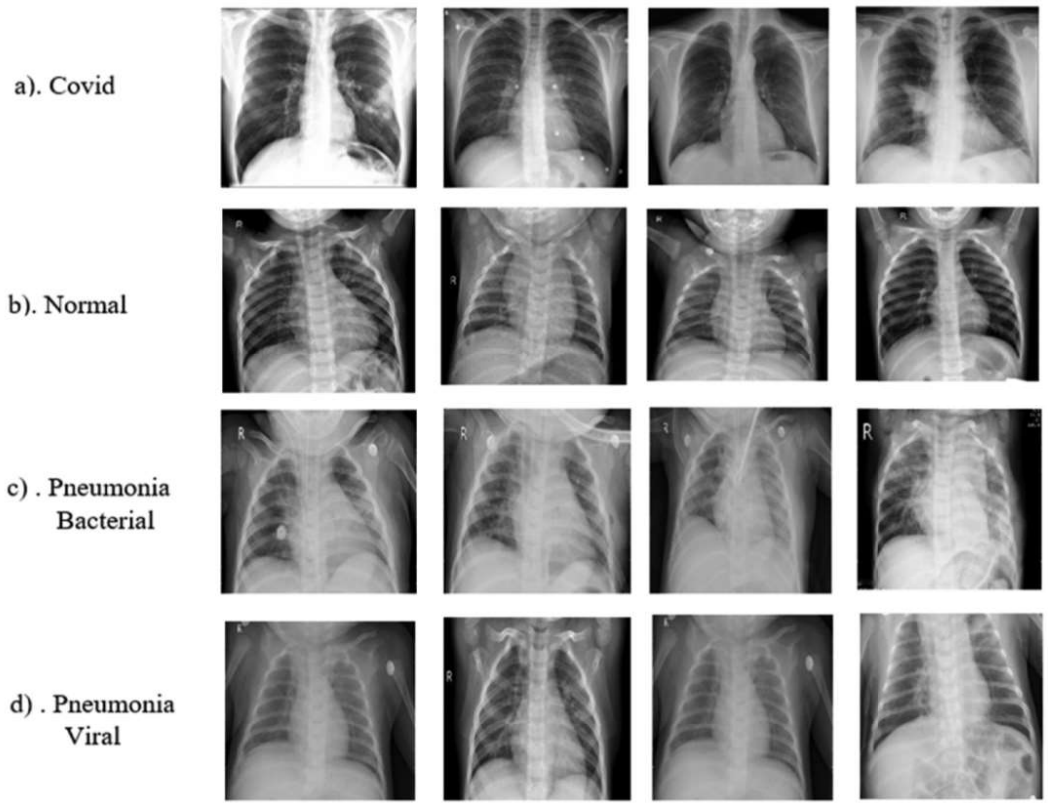

FIGURE 3.2: Chest X-Ray dataset.

i Four Classes – *a. Covid-19*, *b. Normal*, *c. Pneumonia Bacterial (PB)*, *d. Pneumonia Viral (PV).*

ii Three Classes – *a. Covid-19*, *b. Normal*, *c. Pneumonia Bacterial (PB)*

iii Two Classes – *a. Covid-19*, *b. Normal*

3.3.2 Important factors of consideration for Data collection

The inhomogeneities, subject variations, and changes in acquisition parameters—can significantly impact the outcomes of training deep learning models on medical image datasets like ChexNet. Here's an explanation of how each of these factors can affect training outcomes:

1. Contrast Inhomogeneities: Medical images, including X-rays, can exhibit variations in contrast due to factors such as the imaging equipment used, the patient's condition, and the imaging technique. These variations can impact the visual features that the model learns during

training. If not addressed, the model might struggle to generalize across images with differing contrast levels, potentially leading to reduced diagnostic accuracy. Pre-processing techniques that normalize contrast or apply histogram equalization can help mitigate these variations.

2. Subject Variations from Patient to Patient: Patients' physical characteristics, body positions, and underlying medical conditions can result in significant variability in the appearance of pathologies in medical images. Deep learning models trained on a dataset with diverse subject variations might inadvertently focus on irrelevant features or be sensitive to non-essential differences. Robust models need to learn the essential disease-related patterns while ignoring irrelevant variations. Data augmentation techniques, combined with careful curation of diverse patient cases, can help models become more resilient to subject variations.

3. Changes in Acquisition Parameters: Different medical facilities might employ varied acquisition parameters such as X-ray machine settings, exposure levels, and positioning protocols. These parameter differences can introduce image variations that the model may not have encountered during training. Consequently, the model's performance might degrade when applied to images acquired with different parameters. To address this, a diverse dataset that incorporates a range of acquisition settings can aid the model in learning to handle variations effectively.

In addressing these challenges, researchers and authors often take the following steps:

Data Augmentation: By applying various transformations during training, like rotations, flips, and deformations, models can learn to handle subject variations and changes in acquisition parameters.

Normalization: Normalizing pixel values and contrast can help mitigate the effects of contrast inhomogeneities, making the data more consistent for the model.

Data Diversity: Ensuring that the dataset encompasses a broad range of patient profiles, disease severities, and acquisition conditions can improve the model's generalization capabilities.

Regularization Techniques: Techniques like dropout and batch normalization can help models become more robust to variations during training.

Validation and Testing: Rigorous validation on diverse datasets, including data collected from different sources, can reveal how well the model generalizes across variations. Testing the model's performance on external datasets can provide insight into its real-world applicability.

In summary, understanding and addressing contrast inhomogeneities, subject variations, and changes in acquisition parameters are crucial for developing deep learning models on medical image datasets like ChexNet. A combination of pre-processing techniques, diverse data curation, and appropriate training strategies can help models navigate these challenges and produce more reliable and generalizable diagnostic outcomes.

3.3.3 Data Pre-processing

The dataset incorporates X-ray images from diverse age groups, prompting a closer examination of the relationship between age, pathology presentation, and diagnostic accuracy.

1. Age-Related Pathology Variation: Notably, the ChexNet dataset comprises X-ray images from individuals spanning various age groups. This variation in age introduces the possibility of distinct pathology presentations across these groups. Pathologies might manifest differently due to developmental factors, changes in the immune system, or other age-related influences, potentially influencing the accuracy of diagnostic models.

2. Age-Dependent Diagnostic Criteria: Another pertinent consideration is the influence of age on diagnostic criteria. The criteria for identifying specific pathologies could indeed differ across age groups. Physiological and developmental disparities might lead to different diagnostic thresholds. Thus, a model trained predominantly on one age group might not seamlessly apply its learned criteria to others, which is a limitation for a generalized model.

3. Age-Constrained Model Performance: It's worth noting that training models on age-specific subgroups could enhance accuracy and precision for those groups. Similar age cohorts might exhibit consistent pathological patterns that the model can more effectively learn. However, this approach risks creating models that are optimized for narrow age ranges and struggle with unfamiliar age groups due to the complexities tied to age-related variations.

4. Model Generalization Challenge: The cautionary insight about model generalizability is pertinent. If a model trained predominantly on one age demographic is employed to analyze X-ray images from distinct age groups, its performance might diminish. Variances in pathology presentation and diagnostic criteria could compromise its efficacy, highlighting the need for models capable of accommodating diverse age-related factors.

5. Age-Related Pathological Evolution: Indeed, the evolution of pathologies can diverge across different age groups. The progression of diseases might differ in pediatric, adult, or elderly patients, warranting consideration of age-related nuances when constructing and assessing diagnostic models.

To recap, while pursuing heightened accuracy by focusing on specific age groups during ChexNet model training is appealing, it's pivotal to strike a balance between precision and broad applicability. Comprehensive evaluation across diverse age groups and external datasets is essential to ensure models perform adeptly across various cases. Acknowledging and addressing age-associated disparities in pathology presentation, diagnostic criteria, and pathological progression is key to developing resilient and dependable diagnostic models.

Dataset has extensive pre-processing to avoid class imbalance due to insufficient sample images from one particular class. So, to achieve that, we have applied the data augmentation method to increase the number of sample images and make it a uniform distribution. In order to perform data augmentation, we have tried the ImageDataGenerator module from Keras library by tuning various data augmentation parameters like $rescale = 1/255$, $rotation\_range = 20$, $width\_shift\_range = 0.2$, $height\_shift\_range = 0.2$, and, also, we set $horizontal\_flip = True$, as in chest Xray, we are mainly focusing about the vertical area, so horizontal flip does not make any such difference and hence does not lose much information.

In order to remove duplicate images, we searched for exact duplicates or near exact duplicates with the help of Pillow library in python. With the help of hashing, we successfully removed these kinds of duplicates.

We have also selected the same number of images from each class to avoid overfitting or underfitting to one particular class. The network remains unbiased to any specific class when performing classification tasks. Here we use the dataset as data2, data3, and data4, namely for two classes, three classes, and four classes.

3.3.4 Prepared Dataset

The dataset has been divided into three parts for training: Training data, Validation data, and Test data, typically in a ratio of 70% Training data, 20% Validation data, and the rest 10% as Testing data. Testing data remain untouched during the training part. Hence, testing classification accuracy is significantly lower than training and validation classification accuracy, which we will discuss later

in the results section. Here we present Figure.3.3 for the number of images for each class used for training the proposed neural network architecture.

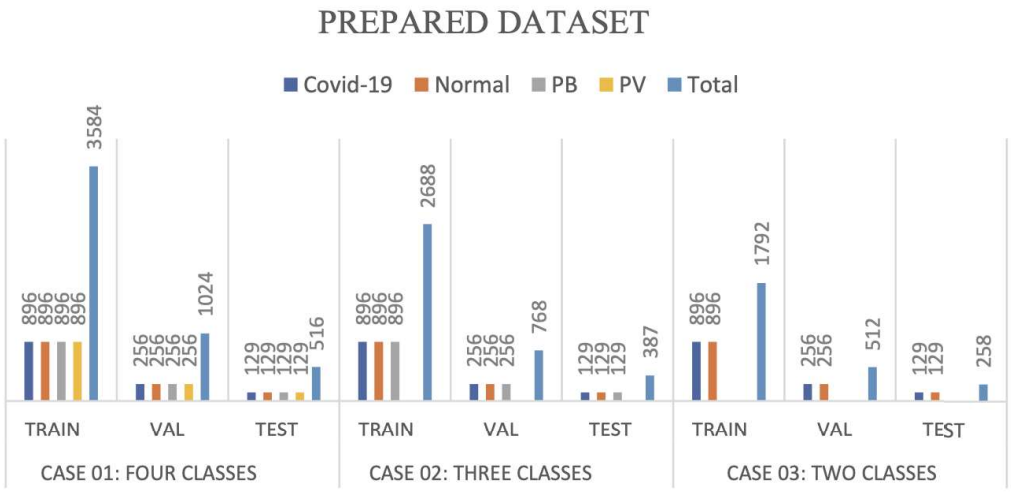

FIGURE 3.3: Prepared Dataset.

3.4 Method and Architecture

3.4.1 Basic block diagram

The basic block diagram of the proposed model architecture operation is given below in Figure.3.4. It starts with data preparation into training (70%), validation (20%) and testing (10%). After that, we performed some data pre-processing and augmentation operations as described in the data pre-processing section. After that, we prepared our transfer learning-based model by freezing all the input layers except the output layer. Then, we modified the output layer as per our desired goal. We then trained our model and evaluated the performance of our model against pre-existing methods.

3.4.2 Inception Module

It is a well-established fact to note that the deeper the network, the better its performance, but it comes with a few drawbacks, such as a more extensive network means a higher number of parameters and suffers from overfitting if the training dataset is minimal. Another major drawback

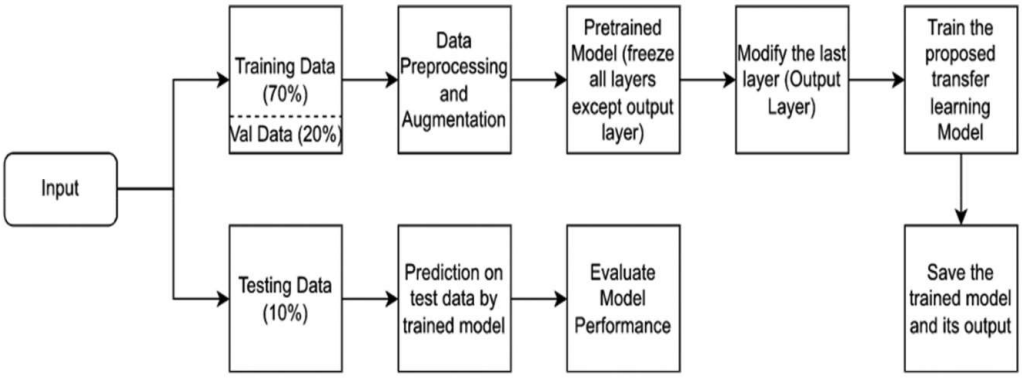

FIGURE 3.4: Basic block diagram of transfer learning-based model.

is an increase in computational complexity and resource requirements. To deal with these problems, we promote sparse architectures rather than dense ones.

It allows the internal layers to pick and choose which filter size will be relevant to learning the required information. A larger filter size (3×3) is used for global feature extraction, and a smaller filter size (2×2) gives the information about distributed features. When the size of the target in the image is different, the layer works accordingly to recognize the target. In the present model, we have used this benefit of the Inception block originally used in Inception v1 to extract features from the images. Once these features are extracted by different filters, they are concatenated before they are fed to the next layer as shown in Figure.3.5.

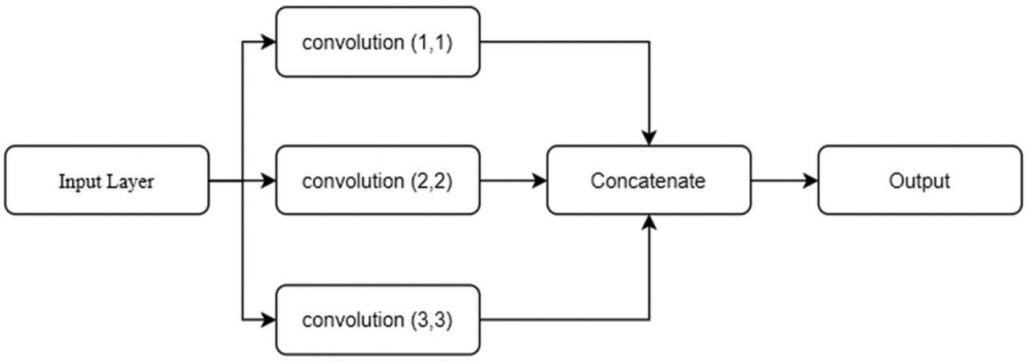

FIGURE 3.5: Inception Module.

3.4.3 VGG-16 Net

The VGG-16 Net is in Figure.3.6. The basic principle is to investigate neural network depth for large-scale visual recognition by utilizing multiple tiny (3×3) convolution filters to increase the network depth. ImageNet Challenge 2014 used VGG-16 Network for large-scale image recognition into 1000 classes. The original VGG-16 net used for this challenge used an image size of 224×224, and for the pre-processing step, they just subtracted its mean value. Then, the image is passed through several (3×3) convolution layers along with stride set to 1, and padding is also set to 1. They used several max-pooling layers following convolution layers to preserve the spatial resolution after convolution. After going through several iterations to these convolutions and max-pooling layers, they used a stack of dense layers which are fully connected, and the final layer was a SoftMax layer that gave classification into 1000 classes.

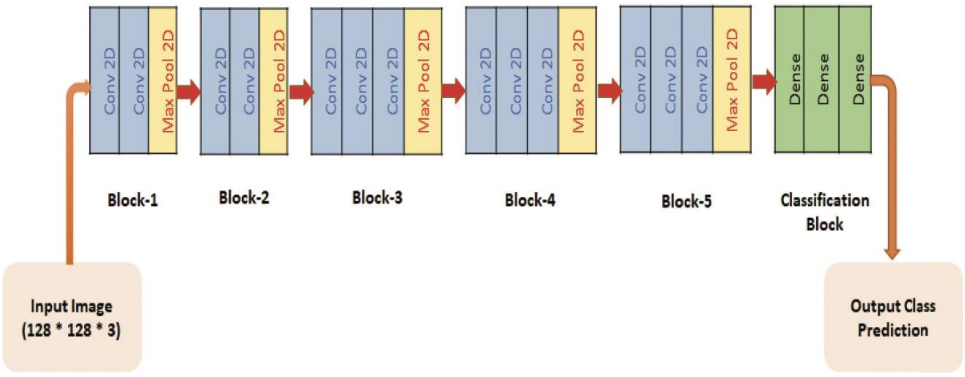

FIGURE 3.6: VGG-16 Net.

3.4.4 Scaled Exponential Linear Units (SELU) activation function

We have seen the popularity of the RELU (Rectified Linear Unit) activation function due to its capabilities to deal with vanishing gradients. This gives zero gradients for negative values and unit gradient as when the input values are positive, defined by Equation.3.1 as-

$$Relu(x) = \begin{cases} 0 & if\ |x| \leq 0 \\ X & if\ |x| > 0 \end{cases} \tag{3.1}$$

But it suffers a dead state when the rate of change of weights is very high, and the resulting value of x is very small in the next iteration that RELU is stuck at the left side of zero. Hence, due to its zero-gradient value, affected cell disconnects to contribute to training in the network. Although you get rid of vanishing gradients, it causes dying RELUs in various cells in the network.

RELU is suitable for its low complexity due to its linear nature for positive x values. Still, to deal with the problem of dying neurons for the negative values, we have modified it to another called Leaky RELU. We now have two options to choose from, either RELU (dying neuron) or Leaky-RELU (little risk of vanishing gradients). But there is another approach, SELU dealing with it with the advantage of its quality to self-normalization.

In this paper, we have used the "SELU" activation function instead of the famous "RELU" function, introduced in September 2017. Basic SELU activation function is defined by Equation.3.2

$$Selu(x) = \lambda \begin{cases} \alpha \exp\{x\} - \alpha & if\ |x| \leq 0 \\ X & if\ |x| > 0 \end{cases} \tag{3.2}$$

It works similarly to RELU for positive values of x, there is no problem with vanishing gradients, and there are no dying neurons for negative values of x. Hence, SELU is our preferred choice for this research work.

3.4.5 Deep Transfer Learning

Since the introduction of CNN, it has found its use in many applications [44], including autonomous car driving, AlphaGo championship, image classification, and segmentation. With the increasing dataset resources and computational capabilities, it has taken rise in its use for the automated diagnosis of diseases using medical images through classification and segmentation. A simple CNN is a sequential stack of multiple layers with many neurons. The deeper the network, with more and more data, the better the predictability of the outcome. However, it has certain limitations [45] as it requires more computational resources and requires an enormous dataset, which might not be available and not in appropriate form. To deal with such a problem, we utilize pre-trained weights and models and change a few output-dense layers according to our need to classify the results. Mathematically [46], we denote it with some notations such as Domain (D) and Task (T).

$$D = \{x, P(X)\} \tag{3.3}$$

where, x–$Feature\ \ Space, P(X)$–$Edge\ \ probability\ \ distribution$, and $X = \{x1, x2, ..., xn\}$

$$T = \{y, f(x)\} \tag{3.4}$$

where, y–$Label\ \ Space, f(x)$–$Target\ \ prediction\ \ function\ \ or\ \ conditional\ \ probability\ \ function$

3.4.6 Proposed Model

The proposed method utilizes the properties of the inception module along with the VGG16 model, which carries three convoluted layers stacked parallel with different kernel window sizes as (1 × 1), (3 × 3), and (5 × 5). These layers have a "SELU" activation function with padding as "same." This module takes four inputs, namely – layer_in, f1, f2, and f3, defined as image input size and sizes of respective kernel filter windows of three convolution layers, respectively. In our experiment, we have taken image input size as (128 × 128 × 3) and f1 = 16, f2 = 16, and f3 = 32.

The modified VGG-16 net in our research work starts with the inception block in which we performed a concatenation of three stacked convolutional layers, whose output will serve as an input to the VGG-16 neural network as shown in Figure.3.7. VGG-16 network consists of five blocks (Block-1 to Block-5) and the final classification block. We start with the first block, which takes input from the output of the inception module. Block-1 consists of two Conv2D layers along with one max pooling layer. Parameters used for these Conv2D layers are number_of_filters = 64, kernel_size = 3, stride = (1,1), padding = same, learning_rate = 0.0002 and for Maxpooling2D layer, pool size was set to (2,2), stride = (2,2) and padding = valid. Block 2 is similar to Block 1 except for the increase in filters from 64 to 128. Block three is similar to the previous block, but with an additional Conv2D layer and increased filters from 128 to 256 for Blocks 4 and 5 with increased filters. Finally, the generated features feed into three sequential dense layers with appropriate activation functions for classification. We used the SoftMax activation function for multi-class classification while using the sigmoid activation function for binary classification. At the last Dense layer, we select the number of neurons as the total number of classes to classify. We have selected 4,3,1 as the number of neurons for four-class, three-class, and binary class classification, respectively, at the output layer.

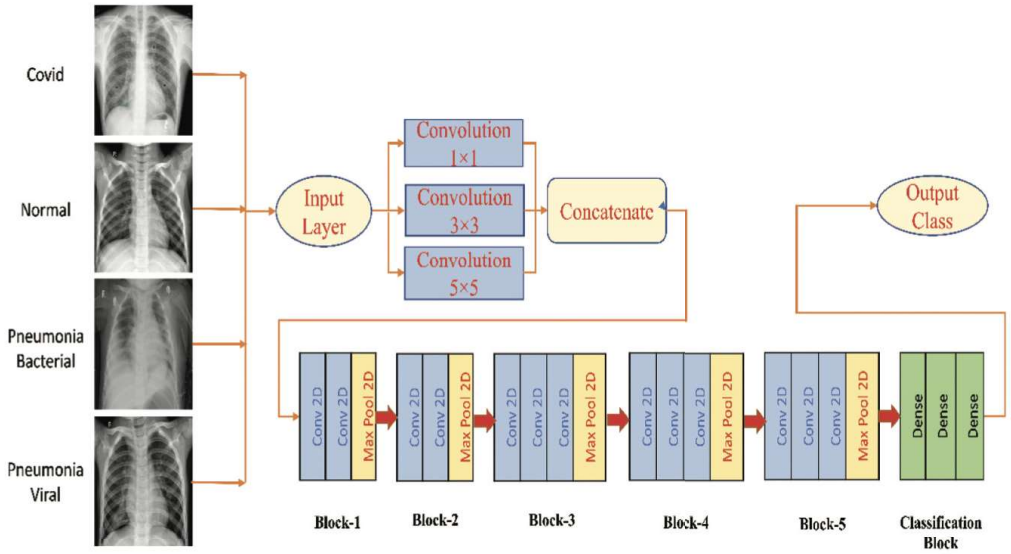

FIGURE 3.7: Proposed Model

3.4.7 Training the Network

In this proposed model, we have utilized pre-trained VGG-16 Net and Inception Net. In order to compare the performance of our model, we have trained our dataset on VGG-16 Net, Res-Net, and Inception-Net. In Table.3.1, we have summarised the performance parameters of original papers with their implementation year and the number of parameters.

Model Parameters				
Network	**Year**	**Salient Feature**	**Top-5 accuracy**	**Parameters**
VGG-16 Net	2014	Fixed-Size Kernels	92.30 %	138M
Inception Net	2014	Wider-Parallel Kernels	93.30 %	6.4M
RES-Net	2015	Shortcut Connections	95.51 %	60.3M
Proposed Net	2022	Fixed-and wider Parallel Kernels	86.32%(Case01-4 Class) 96.89% (Case02-3 Class) 99.98%(Case03-2 Class)	56.7M

TABLE 3.1: Model Parameters

We have utilized the Keras 2.3.1 framework with Tensorflow 2.1.0 and the scikit-learn library to train the proposed model. Complete experiments were performed on a Lenovo Legion Y730 Laptop with configurations such as—Nvidia RTX 2070Q, 8GB DDR5 GPU, 16GB DDR4 RAM, CUDA 11.2 with 1TB SSD. We have used the "SELU" activation function and input image dimension as $128 \times 128 \times 3$. We have trained data on three pre-existing networks by freezing the input layers and then train the network by using "Image-Net" weights. We have similarly trained our model.

Training parameters include, batch_size = 16, steps_per_epoch = 100, and learning_rate = 0.0001 and model was trained for 50 epochs. These model hyperparameter values have been summarised in Table.3.2.

Hyper parameters for training			
Parameter	**Value**	**Parameter**	**Value**
1. train:val:test	70:20:10	6. Number of iterations	50
2. Learning Rate	0.0001	7. Pooling Size	2
3. Optimization algo	ADAM	8. Batch Size	16
4. Activation Function	SELU	9. Kernel Size in CNN	3
5. Cost or loss function	categorical crossentropy, kl_divergence	10. Number of layers	29

TABLE 3.2: Hyper parameters for training

3.4.8 Performance Evaluation Parameters

A classification model is used to find the separate classes as an outcome belonging to a particular category by finding the maximum probability of occurrence. The confusion matrix defines the performance of the classifier system by finding out the cases of True positive (TP), True negative (TN), False positive (FP), False negative (FN), and so on. A typical confusion matrix for Binary class classification is shown below in Figure.3.8 and for Multiclass classification, confusion matrix is shown in Figure.3.9

		PREDICTED CLASS	
		POSITIVE (Covid-19)	**NEGATIVE (Normal)**
ACTUAL CLASS	**POSITIVE (Covid-19)**	True Positive (T.P.)	False Negative (FN.)
	NEGATIVE (Normal)	False Positive (FP.)	True Negative (TN.)

FIGURE 3.8: Binary Class Confusion Matrix

Here, Let's take example of confusion matrix assuming to predict Covid-19 (Covid-19 prediction),then you can evaluate the various performance parameters by finding TP,FP,TN and FN as follows-

		PREDICTED CLASS		
		Covid-19	**Normal**	**Pneumonia**
ACTUAL CLASS	**Covid-19**	Cell (1)	Cell (2)	Cell (3)
	Normal	Cell (4)	Cell (5)	Cell (6)
	Pneumonia	Cell (7)	Cell (8)	Cell (9)

FIGURE 3.9: Multi Class Confusion Matrix

$$TruePositive(TP) = Cell(1)$$
$$FalseNegative(FN) = Cell(2) + Cell(3)$$
$$FalsePositive(FP) = Cell(4) + Cell(7)$$
$$TrueNegative(TN) = Cell(5) + Cell(6) + Cell(8) + Cell(9)$$

All the performance metrics (Fig. 10-15) are utilizing these values to find out metrics like Accuracy, Precision, Sensitivity, Specificity, False-Positive Rate (FPR), F1-Score, and so on, as given below.

1. **Accuracy-** It can be defined by Equation.3.4 and the accuracy parameter for all case studies are shown in Figure.3.10-

$$Accuracy(ACC) = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.5}$$

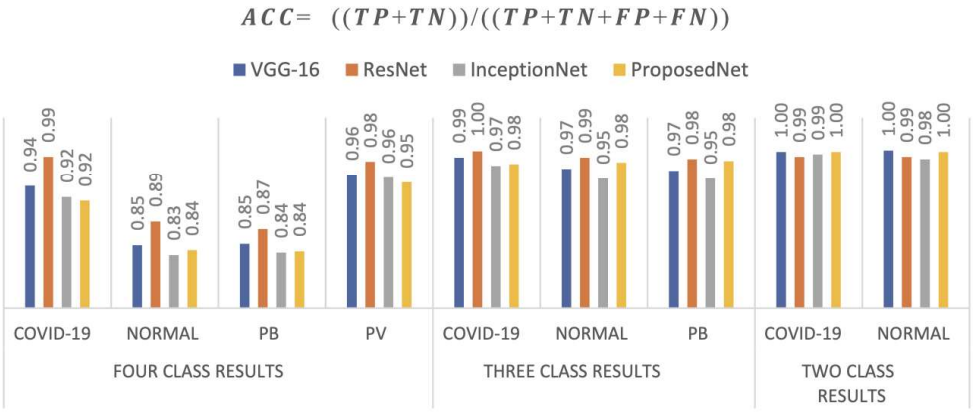

FIGURE 3.10: Accuracy

2. **Precision or positive predictive value (PPV)-** it can be defined by Equation.3.5 and the precision parameter for all case studies are shown in Figure.3.11

$$PR=TP/((TP+FP))$$

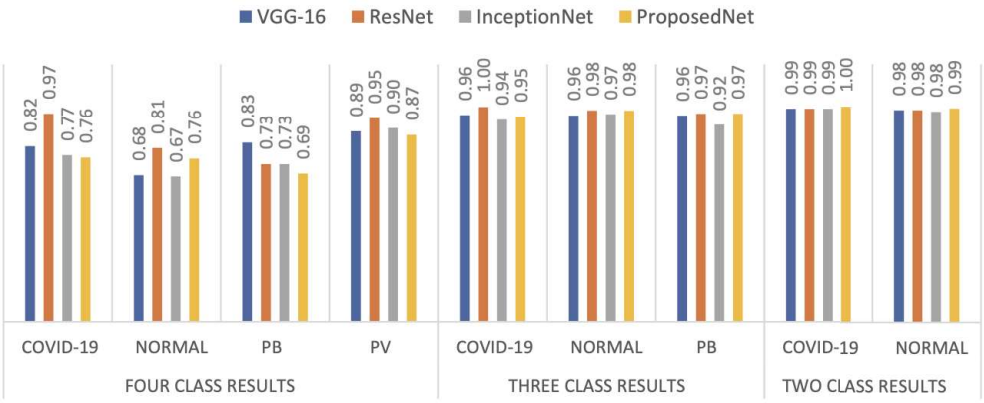

FIGURE 3.11: Precision

$$Precision = \frac{TP}{TP + FP} \tag{3.6}$$

3. **Sensitivity, Recall, Hit rate, or True positive rate (TPR)-** It can be defined by Equation.3.7 and the value of this performance parameter for all case studies are shown in Figure.3.12-

$$Sensitivity(SN) = \frac{TP}{TP + FN} \tag{3.7}$$

4. **Specificity,Selectivity or True negative rate (TNR)-** It can be defined by Equation.3.8 and the value of this performance parameter for all case studies are shown in Figure.3.13

$$Specificity(SP) = \frac{TN}{FP + TN} \tag{3.8}$$

5. **False Positive Rate or Fall-out-** It can be defined by Equation.3.9 and the value of this performance parameter for all case studies are shown in Figure.3.14

$$FalsePositiveRate(FPR) = \frac{FP}{FP + TN} \tag{3.9}$$

$$SN = TP/((TP+FN))$$

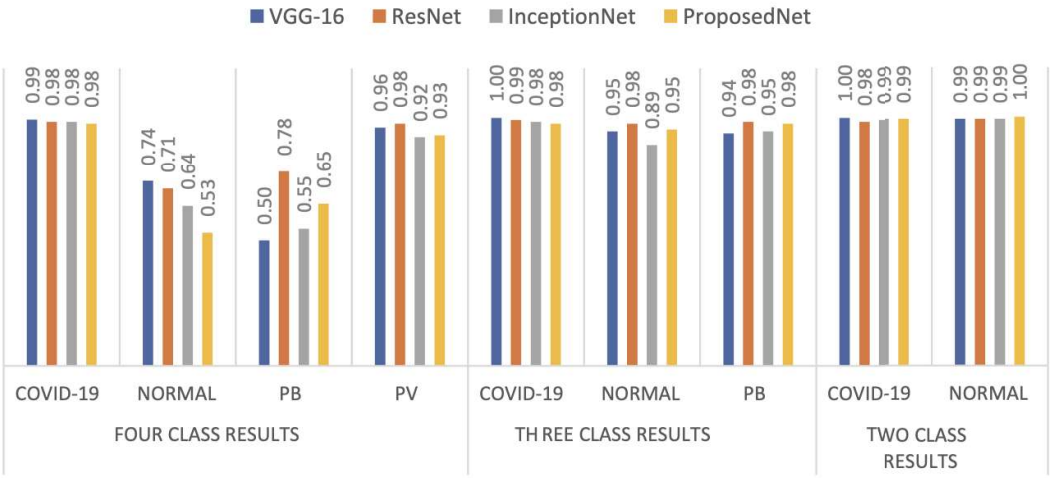

FIGURE 3.12: Sensitivity

$$SP = TN/((FP+TN))$$

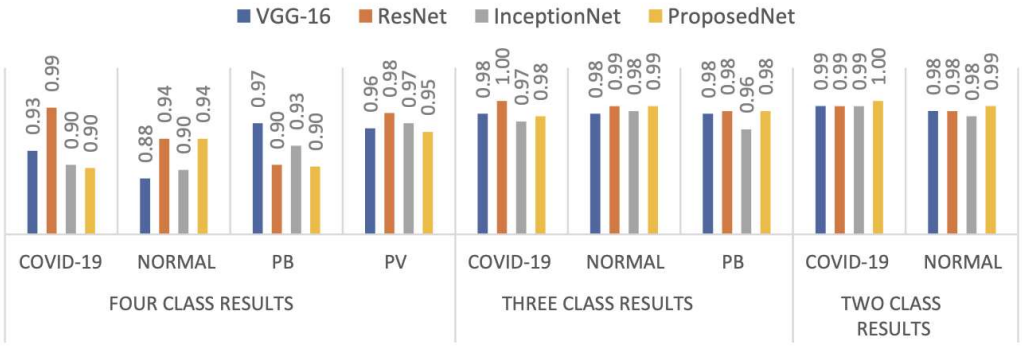

FIGURE 3.13: Specificity

6. **F1-Score-** It can be defined by Equation.3.10 and the value of this performance parameter for all case studies are shown in Figure.3.14

$$F1 - Score = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN} \tag{3.10}$$

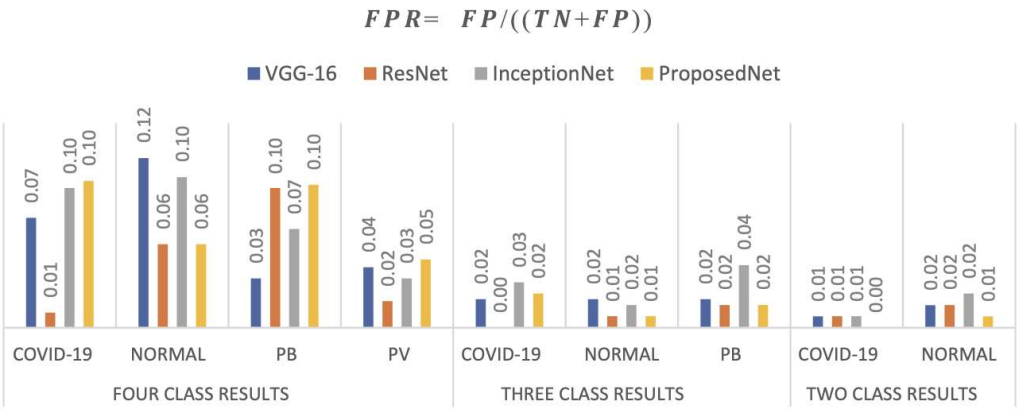

FIGURE 3.14: False Positive Rate (FPR)

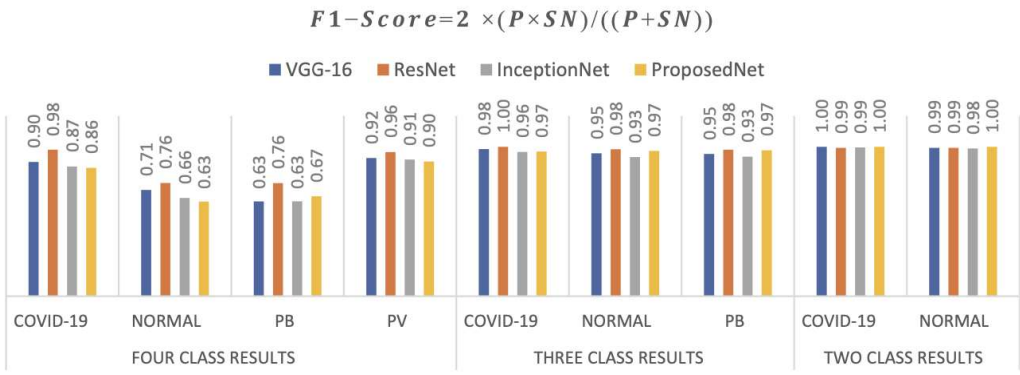

FIGURE 3.15: F1-Score

3.5 Experiments and Results

3.5.1 Overall Model Performance

1. **Case01-Four Class Results**

   Four class classification results displays the model performance parameter to classifiy among-Covid-19, Normal, Viral Pneumonia and Bacterial Pneumonia. It has been shown in Figure.3.16

2. **Case02-Three Class Results**

   Three class classification results displays the model performance parameter to classifiy among-Covid-19, Normal, and Bacterial Pneumonia. It has been shown in Figure.3.17

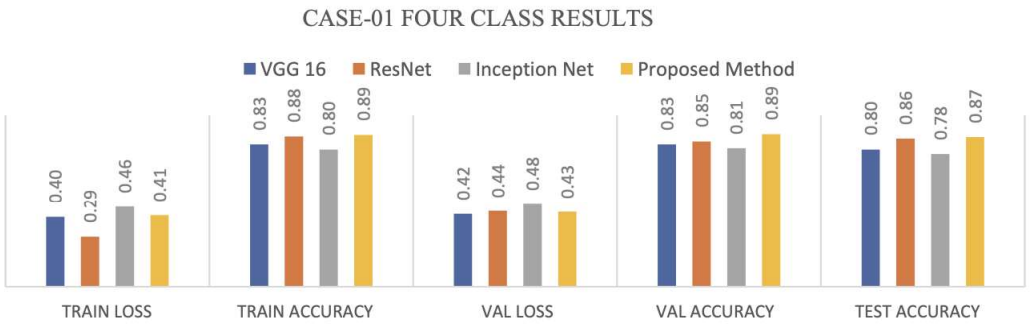

FIGURE 3.16: Case 01-Four Class Results

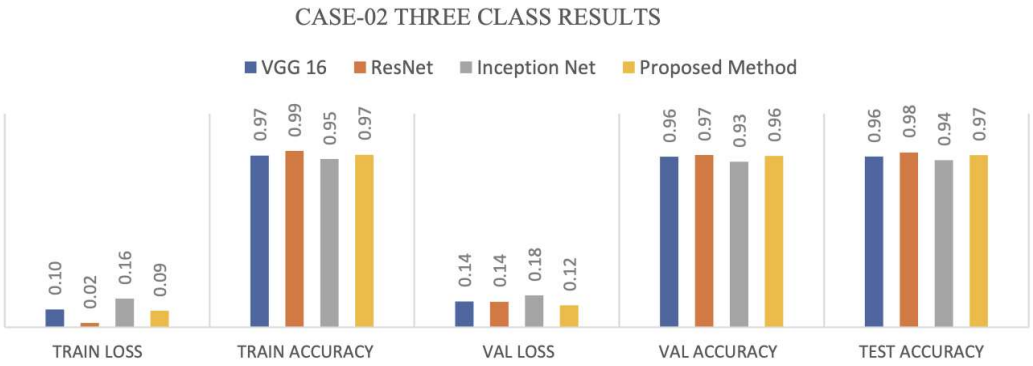

FIGURE 3.17: Case 02-Three Class Results

3. **Case03-Two Class Results**

Two class classification results displays the model performance parameter to classifiy between-Covid-19 vs Normal. These results are shown in Figure.3.18

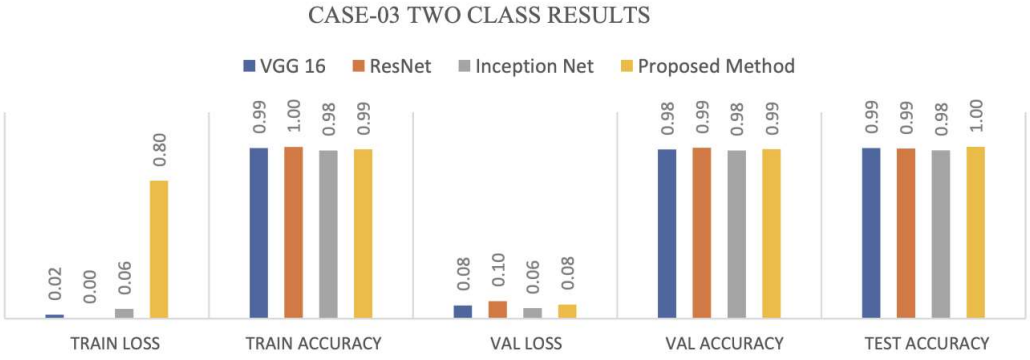

FIGURE 3.18: Case 03-Two Class Results

3.5.2 Performance evaluation parameters for all three cases

Six performance parameters – Accuracy, Precision, Specificity, Sensitivity, False-Positive Rate (FPR), and F1-Score – have been shown in Figure.3.10-3.15. We have calculated these parameters and compared them with three pre-existing models. We have trained and tested this dataset for all models and cases. Figure.3.16-3.18 give us clear insights into which performance metrics performed better in our proposed model. We have seen performance improvement despite considering that we are using a smaller dataset and utilizing less computational power than pre-trained VGG-16 Net, Res-Net, and Inception-Net.

3.5.3 How to improve performance of Model

Improving the performance of a 4-class classification model on the ChexNet dataset, which includes Covid, normal, viral pneumonia, and bacterial pneumonia cases, involves a combination of data-related strategies and model optimization techniques. Here's a comprehensive response on how the model's performance can be enhanced:

1. Data Augmentation:

Expanding the dataset through data augmentation techniques [47] can improve model generalization. By applying transformations like rotations, flips, and scaling to the images, the model learns to recognize the same pathology from different perspectives, making it more robust to variations in presentation.

2. Balancing Class Distribution:

Ensure that the dataset has a balanced representation of all four classes [48]. If one class has significantly fewer samples, the model may struggle to learn patterns from that class. Techniques like oversampling, undersampling, or generating synthetic samples can help achieve a more equitable distribution.

3. Fine-Tuning and Transfer Learning:

Start with a pre-trained model [49], such as one trained on a large dataset like ImageNet. Fine-tune the model's weights using the ChexNet dataset. This leverages the features learned from a larger dataset and can lead to improved performance on the target classes.

4. Architecture Selection:

Experiment with different deep learning architectures[50] suitable for image classification tasks. Popular architectures include Convolutional Neural Networks (CNNs) like ResNet, DenseNet, and Inception. The architecture's depth and complexity should match the size of the dataset and the complexity of the classification task.

5. Hyperparameter Tuning:

Tune hyperparameters like learning rate, batch size, optimizer, and regularization strength. A grid search or random search approach can help identify the optimal combination of hyperparameters that yield the best performance[51].

6. Regularization Techniques:

Implement regularization techniques [52] like dropout and batch normalization to prevent overfitting and improve model generalization to unseen data.

7. Attention Mechanisms:

Utilize attention mechanisms [53], such as self-attention or spatial attention, to enable the model to focus on more informative regions of the images. This can enhance the model's ability to capture subtle patterns.

8. Ensemble Learning:

Combine predictions from multiple models to create an ensemble. Ensemble methods [54] can improve performance by reducing bias and variance and capturing diverse patterns in the data.

9. Progressive Learning:

Train the model progressively[55] by starting with simpler tasks and gradually increasing the complexity. This approach allows the model to learn progressively more challenging patterns, leading to better convergence.

10. Transfer Learning with Domain Adaptation:

If there are domain-specific variations in the ChexNet dataset, consider techniques like domain adaptation [56]. This involves adapting the model to the target domain while leveraging knowledge from the source domain.

11. Interpretable AI Techniques:

Employ techniques like Grad-CAM (Gradient-weighted Class Activation Mapping) to visualize and interpret the areas of the image that the model focuses on when making predictions. This can provide insights into the model's decision-making process [57].

12. Regular Monitoring and Model Updating:

Continuously monitor the model's performance on validation and test sets. If the performance plateaus or degrades, consider updating the model with new data or refining the training process.

In conclusion, improving the performance of a 4-class classification model on the ChexNet dataset requires a multifaceted approach that includes data augmentation, balanced classes, architecture selection, hyperparameter tuning, regularization, and potentially advanced techniques like attention mechanisms and ensembling. Experimentation, iterative refinement, and a deep understanding of the dataset's nuances are key to achieving optimal performance.

3.5.4 Impact of model evolution over time

in the context of medical image analysis like the ChexNet dataset. Here's an explanation of how models' capabilities might evolve with changing diagnostic criteria over time:

1. Dynamic Nature of Diagnostic Criteria: Diagnostic criteria for medical conditions are not static; they evolve as our understanding of diseases improves. As medical research advances, new insights and discoveries lead to refined criteria for diagnosing various pathologies. This dynamic nature reflects the complexity of medical science and the ongoing pursuit of accuracy in diagnosis.

2. Model Adaptation to Changing Criteria: Deep learning models like those used in medical image analysis, including ChexNet, possess the potential to adapt to changing diagnostic criteria. However, this adaptation is contingent upon timely updates to the model's training data and retraining processes. When diagnostic criteria change, incorporating updated data that reflects these changes becomes essential to ensure the model remains accurate and aligned with current medical knowledge.

3. Continuous Learning and Retraining: One way to enable models to accommodate changing diagnostic criteria is through continuous learning and retraining. Models can be periodically retrained using new data that reflects the updated criteria. This process allows the model to learn the evolving patterns and characteristics associated with the revised criteria.

4. Incremental Improvement: Over time, as models are exposed to more varied and up-to-date data, their performance can incrementally improve. This improvement reflects the model's ability to learn and generalize from a broader range of cases. As new data with revised diagnostic criteria is introduced, the model may gradually refine its ability to accurately diagnose diseases based on the latest understanding.

5. Challenges and Caveats: Adapting models to changing criteria comes with challenges. Models need to be validated rigorously after updates to ensure they retain their generalization capabilities and do not overfit to specific changes. Striking a balance between retaining knowledge learned from previous data and incorporating new information is vital to maintaining robust model performance.

6. Collaboration with Medical Experts: Close collaboration between machine learning practitioners and medical experts is essential. Medical professionals can provide insights into changing diagnostic criteria, helping guide the model's retraining process effectively and ensuring it aligns with the current medical landscape.

In summary, while deep learning models have the potential to improve their capabilities with changing diagnostic criteria, this improvement requires a proactive approach involving continuous learning, retraining, validation, and collaboration between machine learning experts and medical professionals. By incorporating updated data and maintaining a flexible learning approach, models can adapt to evolving medical knowledge and contribute to more accurate and relevant diagnoses.

3.5.5 Evaluating model robustness: Age-Group Generalization

To acknowledge the importance of age and gender-related biases in medical imaging, and in order to validate our model on age-related variations [58], I have downloaded a smaller dataset sample from Kaggle, which is specifically mentioned as the children's chest X-ray dataset. The dataset is organized into 3 folders (train, test, val) and contains subfolders for each image category (Pneumonia/Normal). There are 5,863 X-Ray images (JPEG) and 2 categories (Pneumonia/Normal).

Chest X-ray images (anterior-posterior) were selected from retrospective cohorts of pediatric patients of one to five years old from Guangzhou Women and Children's Medical Center, Guangzhou. All chest X-ray imaging was performed as part of patients' routine clinical care.

For the analysis of chest X-ray images, all chest radiographs were initially screened for quality control by removing all low-quality or unreadable scans. The diagnoses for the images were then

graded by two expert physicians before being cleared for training in the AI system. In order to account for any grading errors, the evaluation set was also checked by a third expert.

I took a sample dataset from the above which includes 1400 images, which has been split into training (1120 images) and testing (280 images) sets with a split ratio of 0.75. It consists of three classes, namely - Normal, Viral Pneumonia, and Bacterial Pneumonia. I have trained the proposed model on this dataset, and the model's test performance results are as follows:

Classification Results (Normal Vs Pneumonia for children chest X-Ray dataset)

	Predicted Positive	Predicted Negative
Actual Positive	130	30
Actual Negative	35	85

Accuracy

(TP + TN)/(TP + TN + FP + FN) = = (130 + 85) / (130 + 85 + 35 + 30) = 67.83 %

Presicion

TP / (TP + FP) = 130 / (130 + 35) = 78.79 %

Recall

TP / (TP + FN) = 130 / (130 + 30) = 81.25 %

F1-Score

2 * (Precision * Recall) / (Precision + Recall)

= 2 * (0.7879 * 0.8125) / (0.7879 + 0.8125) = 79.99 %

3.5.6 Model Performance Comparison (Adult vs Children)

Upon scrutinizing the outcomes presented above and in direct contrast to the model's performance when applied to the adult dataset, it is evident that the corresponding metrics exhibit notable disparities [58] [59], [60], [61], [62], [63]. This disparity is visually represented in the accompanying table:

Dataset	Model	Train Accuracy	Val Accuracy	Test Accuracy
Adult Chest X-ray	ResNet	0.99	0.97	0.98
	VGG-16 Net	0.97	0.96	0.96
	Inception Net	0.95	0.93	0.94
	Proposed Net	0.97	0.96	0.97
Children Chest X-ray	ResNet	0.62	0.58	0.61
	VGG-16 Net	0.61	0.59	0.58
	Inception Net	0.59	0.57	0.56
	Proposed Net	0.67	0.61	0.59

In order to Justify poor model performance on a children's X-ray dataset compared to an adult chest X-ray dataset for COVID-19 detection can be attributed to several probable factors. Here is a proper justification for this difference in performance:

1. Age-Related Differences in Disease Presentation:

COVID-19 can manifest differently in children compared to adults. Children often exhibit milder symptoms, and their chest X-rays may not show the same level of pathology as those of adults [61, 62, 64]. This fundamental difference in disease presentation can make it more challenging to detect COVID-19 in children through imaging alone.

2. Dataset Size and Diversity:

The size and diversity of the dataset play a crucial role in model performance. Adult chest X-ray datasets are typically more extensive and well-annotated than children's datasets [65] due to the higher incidence of chest X-ray imaging in adults. A smaller and less diverse dataset for children can limit the model's ability to generalize to a broader population.

3. Data Imbalance:

Children's X-ray datasets may suffer from class imbalance [65], where the number of COVID-19 cases is significantly lower than other conditions or normal cases. This imbalance can lead to bias in the model, making it more proficient at classifying the majority class (e.g., normal cases) and less accurate in identifying the minority class (COVID-19 cases).

4. Developmental Differences in Anatomy:

Children's anatomy differs [7] from that of adults, and their chest X-ray images can vary in terms of size, structure, and growth-related changes. Models trained primarily on adult datasets may struggle to adapt to these age-specific anatomical variances, affecting their performance on pediatric X-rays.

5. Limited Availability of Pediatric Data:

Pediatric-specific COVID-19 datasets, especially those containing a large number of X-ray images, may be limited [66] in comparison to adult datasets. This scarcity of data can hinder model training and result in suboptimal performance.

6. Technical Challenges in Pediatric Imaging:

Capturing high-quality X-ray images of children can be technically challenging[67] due to their smaller size and higher likelihood of movement during the imaging process. Lower image quality can introduce noise and reduce the model's ability to detect subtle abnormalities.

7. Model Optimization for Age Groups:

Models optimized for adult chest X-ray analysis may not be suitable for pediatric cases without fine-tuning or transfer learning [68]. Age-specific model optimization is crucial to adapt to the unique characteristics of pediatric X-rays.

8. Disease Prevalence in the Population:

The prevalence of COVID-19 may differ between adults and children in a given population [69]. A lower prevalence in the pediatric population may result in fewer positive cases for model training, making it more challenging for the model to learn the patterns specific to pediatric COVID-19 cases.

In summary, the poorer model performance on a children's X-ray dataset compared to an adult chest X-ray dataset for COVID-19 detection can be attributed to a combination of differences in disease presentation, dataset characteristics, anatomical variances, and technical challenges. Addressing these challenges may require dedicated efforts, including the collection of more extensive and diverse pediatric datasets and the development of age-specific model optimizations.

3.5.7 Achieving the enhanced generalization capabilities of the model

To develop a model with enhanced generalization capabilities[70] across both adult and pediatric X-ray datasets, while also improving its performance in the context of COVID-19 classification, the following research avenues warrant exploration:

1. Comprehensive Age-Stratified Datasets:

Collect and curate large and comprehensive X-ray datasets that include both adults and children, stratified [71] by age groups. Ensure that these datasets capture a wide range of COVID-19 cases, other pneumonia types, and normal cases for each age group.

2. Anatomical Variation Modeling:

Investigate methods for modeling and accommodating the anatomical variations[72] between adults and children in chest X-ray images. This could involve developing age-specific preprocessing techniques and data augmentation strategies to standardize image size and features.

3. Transfer Learning and Fine-Tuning:

Explore transfer learning techniques by pretraining models on adult chest X-ray datasets and fine-tuning [73] them on pediatric datasets. Fine-tuning should consider the differences in age groups, enabling the model to adapt to both populations effectively.

4. Age and Gender Biases:

Investigate the impact of age and gender biases [61] in COVID-19 classification. Analyze how these biases affect model performance and develop techniques to mitigate bias-related challenges for both adults and children.

5. Ensemble Models:

Evaluate the effectiveness of ensemble models [74] that combine predictions from multiple models, each specialized for a specific age group. This can help improve overall model performance and robustness.

6. Data Augmentation and Synthesis:

Develop advanced data augmentation and synthesis techniques [75] specifically designed for pediatric X-ray images. These methods can help address the challenges posed by smaller datasets for children.

7. Explainable AI in Pediatric Medicine:

Incorporate explainable AI techniques to provide interpretable results for pediatric COVID-19 diagnosis [76]. This can enhance the trust and adoption of AI-based diagnostic tools in pediatric medicine.

8. Clinical Validation and Collaboration:

Collaborate with medical professionals to validate model performance in real clinical settings. Gather feedback from clinicians [77] to fine-tune the models for practical use and ensure alignment with clinical needs.

9. Multimodal Approaches:

Investigate the potential benefits of combining X-ray data with other clinical data modalities [78], such as patient history and laboratory results, to enhance diagnostic accuracy for both adults and children.

10. Robustness to Noise and Image Quality:

Develop models that are robust to variations in image quality and noise [79], which can be common in pediatric X-rays. This can involve preprocessing techniques to enhance image quality.

11. Ethical Considerations:

Address ethical considerations [80] related to pediatric data collection and model deployment, ensuring that privacy and informed consent protocols are in place.

12. Knowledge Sharing and Collaboration:

Foster collaboration among researchers, medical institutions, and data providers to share knowledge, datasets, and best practices for developing generalized models for COVID-19 detection across different age groups.

By considering the above-mentioned points, we can work towards creating more versatile and effective models for COVID-19 detection in both adult and pediatric populations, ultimately improving diagnostic accuracy and benefiting the field of medical imaging and healthcare as a whole.

3.5.8 Discussion and Conclusion

In this paper, we have proposed a transfer-learning-based model, which discusses three case studies

1. Four-Class Classification
2. Three-Class Classification
3. Two-Class Classification

For Case (01), for our proposed model, we have found out the model training and validation accuracy of approx. 88.62% and 89%, respectively, and the testing accuracy on 516 images was approx. 87.32%. We also evaluated other performance parameters for each class as Accuracies – Covid-19 (91.86%), Normal (84.11%), Pneumonia Bacterial (83.91%), and Pneumonia Viral (94.77%).

For Case (02), for our proposed model, we have found out the model training and validation accuracy of approx. 97.13% and 96.48%, respectively, and the testing accuracy on 387 images was approx. 96.89%. We also evaluated other performance parameters for each class as Accuracies – Covid-19 (97.67%), Normal (97.93%), and Pneumonia Bacterial (98.19%).

For Case (03), for our proposed model, we have found out the model training and validation accuracy approx. of 98.50% and 98.63%, respectively, and the testing accuracy on 258 images was approx. 99.95%. We also evaluated other performance parameters for each class as Accuracies – Covid-19 (99.61%) and Normal (99.61%).

All these performance parameters have been evealuated with the confusion matrices as shown in Figure.3.8 and Figure.3.9. After evaluating the performance parameters, we have seen overall model performance comparison in Figure.3.16-3.18.We see that our model performed well in most cases and worked well to identify individual classes. The introduction of the **SELU** activation function enabled us to deal with the vanishing gradients and tackle the problem of dying neurons, which might affect the network training.

Although this method has found promising results, there is still scope for performance improvement. With the availability of larger datasets and computational power, this seems achievable. But still, the purpose of this study was to assist in the early diagnosis of Covid-19 from other flu-like symptoms like Viral Pneumonia and Bacterial Pneumonia. An early diagnosis might be helpful to cluster people into several classes and treat them accordingly. In the future, we can collect more data and try experimenting with more and more models to get better accuracy, which can be helpful for the complete automated diagnosis of Covid-19, just by an X-Ray or CT.