
# COVID-19 Detection from Chest X-Ray Images

## using Inception V3

MA 305 - Data Science

**Submitted to:**

Prof. M. Tanveer

**Team Members:**

Anmol Jain (230008009)

Nidarsana M (230004031)

Salaj Bansal (230002063)

# Contents

# 1 Abstract

This project develops a deep learning–based model to detect COVID-19 infections from chest X-ray (CXR) images using transfer learning with the **InceptionV3** architecture. The model distinguishes between COVID-19 and Non-COVID cases through efficient preprocessing, data augmentation, and multi-phase fine-tuning, achieving high diagnostic accuracy. The implementation is available at:

**Google Colab Notebook:** `Click here to view the implementation`

# 2 Introduction

The COVID-19 pandemic has emphasized the need for fast and accurate diagnostic tools. Although RT-PCR testing remains the gold standard, it is both expensive and time-consuming. Chest X-rays (CXR), on the other hand, offer a rapid and inexpensive means for preliminary screening.

Radiologists identify COVID-19 based on specific visual cues such as **ground-glass opacities**, **bilateral infiltrates**, and **lung consolidation**. However, manual interpretation is subjective and limited by expertise availability.

Deep learning, specifically Convolutional Neural Networks (CNNs), can automate this process by learning discriminative patterns from radiographs. In this project, we employ the **InceptionV3** architecture, a powerful CNN pretrained on ImageNet, to classify X-ray images into COVID-19 and Non-COVID categories.

The primary objectives of this project are:

- To preprocess and enhance chest X-ray images for consistent input quality.
- To train and fine-tune the InceptionV3 model for accurate COVID-19 detection.
- To evaluate the trained model using accuracy, precision, recall, and AUC.

# 3 Dataset Details

## Dataset Source

The dataset used for this project is the **COVID-19 Chest X-Ray Dataset**, curated by Joseph Paul Cohen and collaborators. It is publicly available at `https://github.com/ieee8023/covid-chestxray-dataset`. This repository compiles chest radiographs from multiple hospitals, research studies, and publications during the COVID-19 pandemic. It includes both the raw image files and a corresponding metadata file `metadata.csv`, which provides rich contextual and clinical information for each image.

### Metadata Overview

The `metadata.csv` file includes the following columns:

`patientid, offset, sex, age, finding, RT_PCR_positive, survival, intubated, intubation_present, went_icu, in_icu, needed_supplemental_O2, extubated, temperature, pO2_saturation, leukocyte_count, neutrophil_count, lymphocyte_count, view, modality, date, location, folder, filename, doi, url, license, clinical_notes, other_notes.`

Among these, the key attributes used in this project are:

- **Finding:** Describes the radiological diagnosis or observation, such as *Pneumonia/Viral/COVID-19*, *ARDS*, or *Normal*.
- **Modality:** Indicates the imaging type. For this project, only *X-ray* images were considered, while CT scans and other modalities were excluded.
- **View Position:** Specifies the projection view — typically *PA (Posteroanterior)*, *AP (Anteroposterior)*, or *Lateral (L)*. Only PA and AP views were retained to maintain anatomical consistency.

## 4 InceptionV3 Architecture Overview

InceptionV3 is a deep convolutional neural network architecture introduced by Szegedy et al. (2016) as an improvement over GoogLeNet. It achieves high accuracy with lower computational cost through efficient architectural innovations.

### Key Design Principles

- **Factorized Convolutions:** Large convolutional filters (e.g., 5×5) are decomposed into smaller ones (e.g., 3×3 and 3x3) to reduce computational complexity.
- **Asymmetric Convolutions:** Instead of a single 7×7 filter, InceptionV3 uses consecutive 1×7 and 7×1 convolutions to achieve the same receptive field at a lower cost.
- **Dimensionality Reduction:** 1×1 convolutions are used to reduce the number of channels before expensive spatial convolutions.
- **Auxiliary Classifiers:** Intermediate branches provide additional gradient flow during training to mitigate vanishing gradients.
- **Batch Normalization:** Applied extensively for faster and more stable convergence.

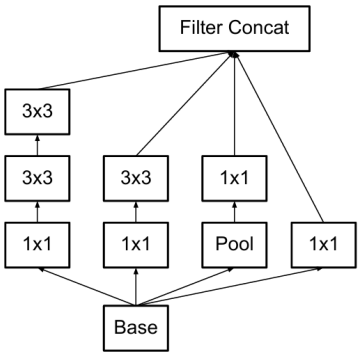


Figure 1: Inception modules where each 5x5 convolution is replaced by two 3x3 convolutions

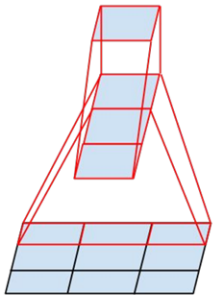

Figure 2: Asymmetric Convolutions - The lower layer of this network consists of a 3x1 convolution with 3 output units

### Structure of our model

This model is built upon InceptionV3. It has been enhanced for COVID-19 chest X-ray classification, combining the powerful base of InceptionV3 with advanced preprocessing, dual pooling, and a multi-layer dense classifier.

# 5 Methodology

This section outlines the complete experimental workflow for COVID-19 detection using InceptionV3. It includes data preprocessing techniques, model design enhancements, a multi-phase training strategy, and the use of Test-Time Augmentation (TTA) for stable predictions.

### Data Loading and Filtering

The following preprocessing and filtering operations were performed sequentially:

- Removed entries with missing values in `finding`, `modality`, or `view`, resulting in a total of **950 valid records**.
- Excluded samples with `view = Lateral (L)` or `view = Axial` to maintain consistent chest orientation. This filtering removed **84 Lateral** and **68 Axial** images, leaving **798 images**.
- Removed **16 CT scan** images (non–X-ray modality), leaving a final count of **782 valid chest X-ray images**.
- Assigned binary classification labels based on the `finding` field:

- **COVID-19:** Entries where `finding` contained "Pneumonia/Viral/COVID-19" (total of **478 images**).
- **Non-COVID:** All remaining chest X-rays (total of **304 images**).

A Pandas DataFrame was then constructed containing two columns: `filepaths` and `labels`, mapping each image file to its corresponding class label.

### Data Balancing and Splitting

Initially, the dataset was imbalanced, with significantly more COVID-19 samples than Non-COVID samples. To address this, the Non-COVID class (**304 samples**) was oversampled through random duplication to match the number of COVID-19 samples (**478**). The resulting balanced dataset contained a total of **956 X-ray images** — evenly split between the two classes.

The dataset was then randomly shuffled and stratified into three subsets:

- **Training Set:** 80% (**764 images**)
- **Validation Set:** 10% (**96 images**)
- **Test Set:** 10% (**96 images**)

This stratified partitioning ensured that both COVID-19 and Non-COVID images were equally represented across all splits, maintaining a balanced class distribution during training and evaluation.

### Data Augmentation and Normalization

To prevent overfitting and increase the model's ability to generalize, extensive augmentation was applied to the training set using TensorFlow's `ImageDataGenerator`. The transformations included:

- Random rotations up to $\pm 30°$.
- Width and height shifts up to 25%.
- Zoom range between 0.75 and 1.25.
- Brightness variation between 0.6 and 1.4.
- Shear transformations up to 0.2 radians.
- Random horizontal flips.
- Channel shift up to 25 units.

The validation and test sets were not augmented to maintain consistent evaluation metrics. All images were preprocessed using the function

`tf.keras.applications.inception_v3.preprocess_input()`,

which rescales pixel intensities to the range $[-1, 1]$ as expected by the InceptionV3 model.

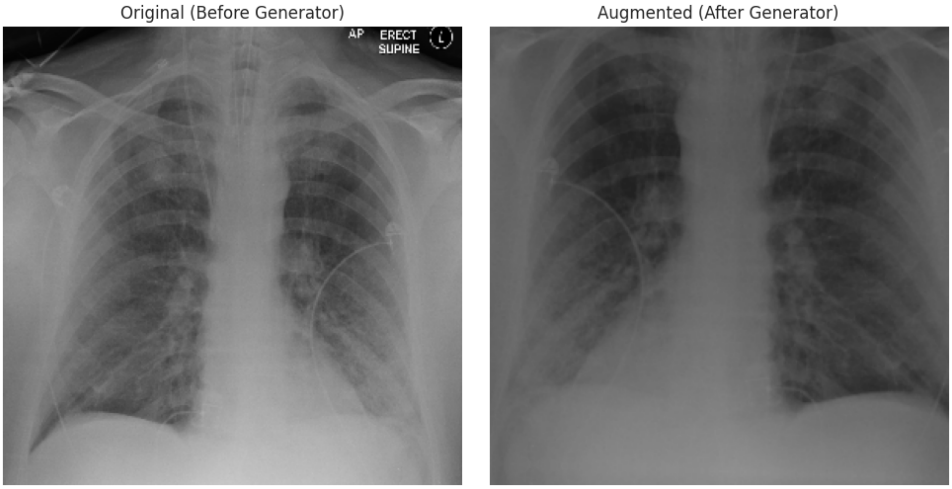


Figure 3: Image Before and After Augmentation

### Data Generators

Custom generators were created for training, validation, and testing. Each generator:

- Loaded images in RGB format at a target resolution of 299×299.
- Returned batches of size 12.
- Encoded labels in categorical format for binary classification.

### Class Weights

To further handle any residual imbalance, class weights were computed using Scikit-learn's `compute_class_weight()` function. This ensured that both classes contributed equally to the loss function during training, preventing bias toward the majority class.

Overall, the data preprocessing stage transformed raw medical images into a clean, balanced, and augmented dataset, ready for model training under the InceptionV3 pipeline.

## Model Design

The core of this project's architecture is an enhanced version of **InceptionV3**, modified for binary classification between COVID-19 and Non-COVID chest X-ray images. The model leverages the pretrained InceptionV3 convolutional base for feature extraction and

introduces a custom classification head with dual pooling and multiple dense layers for improved discrimination capability.

### Base Model

The `InceptionV3` network pretrained on ImageNet was used as the feature extraction backbone. The top fully connected layers were excluded (`include_top = False`) to allow the addition of a task-specific classification head. The pretrained layers were initially frozen to retain their learned feature representations during the early stages of training.

### Dual Pooling Mechanism

To capture both global and localized information from feature maps, a dual pooling strategy was implemented:

- **Global Average Pooling (GAP):** Aggregates the average activation of each feature map, providing a general spatial summary.
- **Global Max Pooling (GMP):** Captures the most activated regions across each feature map, emphasizing key discriminative features.

The outputs of GAP and GMP were concatenated into a single 4096-dimensional feature vector. This combination of average and max activations ensures the model preserves both contextual and salient spatial information—critical for distinguishing infection patterns in X-ray imagery.

### Custom Classification Head

A deep fully connected block was built on top of the concatenated pooled features. Each dense layer was followed by Batch Normalization, ReLU activation, and Dropout for regularization. L1–L2 regularization was applied to reduce overfitting and encourage weight sparsity. The architecture of the classification head is summarized in Table 1.

| Layer Name | Units | Activation | Regularization | Dropout Rate |
|---|---|---|---|---|
| `fc1` | 1536 | ReLU | L1–L2 | 0.6 |
| `fc2` | 768 | ReLU | L1–L2 | 0.5 |
| `fc3` | 384 | ReLU | L1–L2 | 0.4 |
| `fc4` | 192 | ReLU | L1–L2 | 0.3 |

Table 1: Structure of the Custom Dense Classification Head

The final output layer consisted of two neurons with a `softmax` activation function, producing probability scores for each class (COVID-19 and Non-COVID). The output was set to `float32` precision for compatibility with mixed-precision training.

### Regularization and Normalization Techniques

The model incorporated several strategies to enhance stability and prevent overfitting:

- **Batch Normalization:** Applied after each dense layer to stabilize activations and accelerate convergence.
- **Dropout:** Introduced progressively decreasing dropout rates (0.6 to 0.3) to prevent co-adaptation of neurons.
- **L1–L2 Regularization:** Used in all dense layers to penalize large weights and encourage sparse feature representations.

### Model Output

The model's final architecture can be represented as:

$$\text{Output} = \text{Softmax}(W_4(\text{Dropout}_4(\text{ReLU}(W_3(\cdots \text{Concat}(\text{GAP}, \text{GMP})))))))$$

This enhanced InceptionV3 model forms the foundation of the classification system used in subsequent training and fine-tuning stages.

## Training Strategy

The model training was executed in multiple progressive phases to ensure stable convergence and efficient transfer learning. The approach began by training the newly added classification head while freezing the InceptionV3 base, followed by gradually unfreezing deeper convolutional layers with decreasing learning rates. This stepwise fine-tuning strategy allowed the network to adapt pretrained ImageNet weights to the domain of medical X-ray imaging without catastrophic forgetting.

### Phase 1: Initial Training (Frozen Base)

In the initial phase, all layers of the InceptionV3 base model were frozen, and only the custom dense classification head was trained. This step enabled the top layers to learn task-specific decision boundaries while preserving the robust feature extraction capabilities of the pretrained convolutional base. The model was trained for **30 epochs** using the `Nadam` optimizer with a learning rate of **0.001**.

### Phase 2: Fine-Tuning Top 40% of Layers

After initial convergence, the top 40% of the InceptionV3 layers were unfrozen for selective fine-tuning. The remaining 60% of layers stayed frozen to retain generic low-level features such as edges and textures. This phase was trained for **25 epochs** using a reduced

learning rate of **0.0001**. Fine-tuning this limited subset allowed gradual adaptation to the X-ray image distribution without overfitting.

### Phase 3: Fine-Tuning Top 60% of Layers

In the next stage, 60% of the layers were unfrozen, enabling the network to refine both mid-level and high-level features. A lower learning rate of **0.00005** was used to ensure smooth convergence and prevent instability in the pretrained weights. This phase ran for **20 epochs**, further improving feature sensitivity to subtle radiographic differences between COVID-19 and Non-COVID cases.

### Phase 4: Full Model Fine-Tuning

In the final training phase, **all layers** of the InceptionV3 base were unfrozen. The entire model, including both the base and classification head, was fine-tuned end-to-end with an extremely low learning rate of **0.00001** for **15 epochs**. This phase allowed the network to fully adapt to domain-specific visual features while maintaining stability.

### Optimization and Callbacks

To optimize training efficiency and stability, several callback mechanisms were employed:

- **EarlyStopping:** Halted training when the validation F1-score plateaued for 15 epochs, restoring the best weights.
- **ReduceLROnPlateau:** Automatically decreased the learning rate by a factor of 0.2 if validation loss stagnated for 6 epochs.
- **ModelCheckpoint:** Saved the model whenever a new best validation F1-score was achieved.
- **F1ScoreCallback:** Custom callback that computed the weighted F1-score on the validation set after every epoch.

### Performance Monitoring

At the end of each phase, the model's training and validation curves were plotted to monitor the evolution of accuracy and loss across epochs. Figures (4)–(7) illustrate these results.


Figure 4: Training vs Validation Performance – Phase 1 (Frozen Base)


Figure 5: Training vs Validation Performance – Phase 2 (Fine-tuning Top 40%)


Figure 6: Training vs Validation Performance – Phase 3 (Fine-tuning Top 60%)


Figure 7: Training vs Validation Performance – Phase 4 (Full Fine-tuning)

Overall, this phased fine-tuning strategy achieved a balance between preserving the pretrained InceptionV3 features and adapting them to the specific patterns present in chest X-ray images, leading to strong generalization on unseen data.

### Test-Time Augmentation (TTA)

To improve prediction stability, **Test-Time Augmentation (TTA)** was implemented during inference. For each test image, multiple augmented versions were generated by applying small transformations such as:

- Horizontal flip
- Brightness adjustment ($\pm 0.1$)
- Contrast scaling (0.9 and 1.1)

Each augmented image was passed through the trained model, and all resulting softmax probabilities were averaged to obtain the final prediction. This averaging reduces prediction variance and improves overall accuracy and F1-score on unseen data.

## 6 Results and Evaluation

### Model Evaluation

After completing the four-phase fine-tuning process, the best-performing model was saved and reloaded for evaluation on the unseen test set. The evaluation was performed using two approaches:

- **Standard Evaluation:** Direct predictions on test images without augmentation.
- **Test-Time Augmentation (TTA):** Averaged predictions from seven augmented versions of each image for enhanced stability.

## Standard Evaluation Results

On the test set of 96 images, the model demonstrated high generalization performance with the following metrics:

**Test Accuracy:** 88.54% **Test AUC:** 0.9696
**Precision:** 0.8854 **Recall:** 0.8854
**Loss:** 1.67

## Test-Time Augmentation (TTA) Results

Applying TTA further stabilized predictions and marginally improved class-wise balance. It also reduced sensitivity to lighting and orientation variations, enhancing robustness.

**TTA Accuracy:** 88.54% **TTA F1-Score:** 0.8851

## Classification Report

```
============================================================
Classification Report:
============================================================
              precision    recall  f1-score   support

    COVID-19     0.9302    0.8333    0.8791        48
   Non-Covid     0.8491    0.9375    0.8911        48

    accuracy                         0.8854        96
   macro avg     0.8896    0.8854    0.8851        96
weighted avg     0.8896    0.8854    0.8851        96
```

## Confusion Matrix

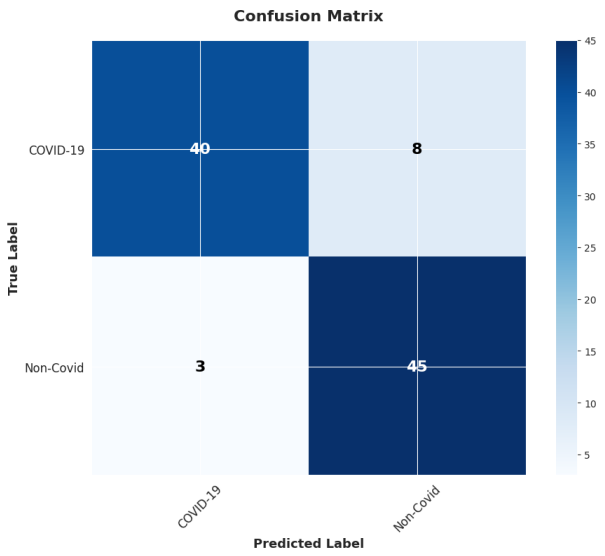


Figure 8: Confusion Matrix for Standard Predictions on Test Set.

### Performance Summary

The model achieved strong and consistent performance across all metrics, confirming the effectiveness of transfer learning with InceptionV3 for COVID-19 detection from chest X-rays. It showed balanced precision and recall across both classes, indicating reliable predictions for both positive and negative cases.
The key evaluation metrics used are defined as follows:

- **Accuracy** $= \dfrac{TP + TN}{TP + TN + FP + FN}$
- **Precision** $= \dfrac{TP}{TP + FP}$
- **Recall** $= \dfrac{TP}{TP + FN}$
- **F1-Score** $= 2 \times \dfrac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$
- **AUC** $= \displaystyle\int_0^1 TPR(FPR^{-1}(x))\, dx$
- **Loss:** Computed using categorical cross-entropy to quantify the prediction error.

$$\text{Loss} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} y_{i,c} \log(\hat{y}_{i,c}) \tag{1}$$

where: $y_{i,c}$ = True label (one-hot encoded), $\hat{y}_{i,c}$ = Predicted probability.

# 7 Discussion

The InceptionV3-based model showed good accuracy and AUC, proving that transfer learning can work effectively for medical images. Staged fine-tuning helped the model adapt gradually without losing previously learned features. TTA improved the stability of predictions by combining results from multiple augmented versions of each image.

However, the main limitation of this study is the limited amount of available data. The dataset not only contains a small number of samples, but it is also imbalanced, with fewer COVID-19 cases compared to Non-COVID images. This imbalance can cause the model to become biased toward the majority class, affecting its ability to detect positive COVID-19 cases accurately. The small dataset size also increases the risk of overfitting and limits the model's ability to generalize to unseen data. Also, variations in image quality and different data sources reduce consistency. Collecting a larger and more balanced dataset would help to improve the accuracy and reliability of the model.

# 8 Conclusion

This implemented an enhanced InceptionV3 model for COVID-19 detection from chest X-ray images. The approach achieved good accuracy and stability through dual pooling and phased fine-tuning. Overall, the model proved that transfer learning can be effectively applied to medical imaging tasks. Future work will focus on expanding the dataset, handling class imbalance, and adding explainability methods for better interpretation of results.

# 9 References


1. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2015). *Rethinking the Inception Architecture for Computer Vision.* arXiv preprint arXiv:1512.00567

2. Sentdex (Harrison Kinsley). (2020). *InceptionV3 Transfer Learning Tutorial.* Link

3. Codebasics (Dhaval Patel). (2021). *TensorFlow Image Classification with InceptionV3.* Video link

4. Agrevolution. (2022). *Test-Time Augmentation for Improving Prediction Accuracy While Inferencing ML Models.* Agrevolution Blog

5. Gupta, M. (2021). *Improving the Accuracy of Image Classifiers.* Medium Blog