Exploring Multimodal Deep Learning: Comparing Pre-trained and Custom Models for COVID-19 Classification

Kazeem Oyebode [1,2], Ebenezer Esenogho [2,*], and Modisane Cameron [2]

[1] Department of Computing, School of Science and Technology, Pan-Atlantic University, Lagos, Nigeria
[2] Centre for Artificial Intelligence and Multidisciplinary Innovation Studies, Department of Auditing, College of Accounting Science, University of South Africa, Pretoria, South Africa
Email: kazeemkz@gmail.com (K.O.); esenoe@unisa.ac.za (E.E.); modistc@unisa.ac.za (M.C.)
*Corresponding author

***Abstract*—COVID-19, a respiratory illness that mostly attacks the human lungs, emerged in 2019 and quickly became a global health crisis. Its fast transmission has necessitated the creation of effective tools that could aid in its classification. In this paper, we present an artificial intelligence multimodal deep learning model that leverages X-ray, Computed Tomography (CT) scan, and cough signals to classify COVID-19 accurately. The paper's objective is to meticulously compare the effectiveness of non-pre-trained and pre-trained versions of VGG-19, MobileNetV2, and ResNet across various multimodal and some unimodal models using cough sound, X-ray, and CT scan datasets. This is important because it provides a pointer as to which combinations of datasets could improve COVID-19 prediction. Findings show that while the pre-trained unimodal systems for cough and X-ray outperform their non-pre-trained counterparts, the non-pre-trained CT scan model performs exceptionally well. This suggests that features learned from the VGG-19 model fail to generalize effectively. Remarkably, the non-pre-trained multimodal model accomplishes an F1-Score of 0.9804, slightly outperforming its pre-trained counterpart at 0.98. While this research advances our understanding of transfer learning, it also emphasizes the prospects of determining, from a range of options, which of the considered datasets (individual or combination) could give an acceptable level of COVID-19 classification in a resource-constrained scenario.**

***Keywords*—machine learning, audio signal processing, deep learning, image classification, multimodal systems, transfer learning, unimodal systems**

I. Introduction

The COVID-19 (Coronavirus Disease 2019) pandemic resulted from the novel SARS-CoV-2 virus that emerged in late 2019 [1]. Its emergence was accompanied by a record number of deaths across the globe. Because of the loss of lives, lockdowns were imminent, leading to businesses shutting down for months to contain the virus. The pandemic strained the healthcare sector to an unprecedented level, as recorded in Europe and the Americas, for example. The World Health Organization has recorded about 2452 COVID-19 deaths from early November 2024 to early December 2024 [2]. This development points to the fact that there is a need for a robust diagnostic system for the early detection of the virus. Its early detection would allow medical personnel to effectively contain it because affected persons would be isolated, contact tracing would start in earnest, and patients would be promptly treated. Furthermore, early detection could potentially save lives [3]. The traditional approach for testing patients for COVID-19 has been through the Reverse Transcription Polymerase Chain Reaction (RT-PCR). This test is usually time-consuming; in addition, it requires specialized skills to carry out the test on patients [4]. This development, therefore, requires that experts be trained on how to effectively administer this process. Also, the process gives many conflicting results, potentially allowing COVID-19 patients to go home due to error [5]. In addition, there have been cases of false positives recorded via RT-PCR [6].

This research aim therefore is to investigate the most promising approach, leveraging artificial intelligence, that could be effective in the prediction of COVID-19 based on available datasets. Datasets considered are X-ray, CT scan, as well as cough sounds.

In response to this, researchers have come up with Artificial Intelligence (AI)-based solutions for the detection of COVID-19 from X-rays [6], Computed Tomography (CT) scans [7], and cough sounds [8]. This is now possible because of the availability of datasets. Cough sound has been used to detect COVID-19, as seen in Ref. [9]. This was combined with patient symptoms to make a prediction. The use of a cough signal is useful because a COVID-19 cough produces a distinctive pattern, differentiating it from a normal cough. This distinctive pattern arises as the virus attacks the lungs, damaging the lung structure [9].

Manuscript received December 30, 2024; revised January 7, 2025; accepted February 17, 2025; published October 24, 2025.

The summary of our key contributions is as follows: To compare the prediction accuracy of COVID-19 using unimodal systems on individual datasets such as CT scan, cough sound, and X-ray with the prediction of COVID-19 using a combination (multimodal) of all the considered datasets. This is important as it provides an avenue for medical personnel and researchers to choose from an array of evidence-based options that best aid objective COVID-19 diagnosis. This work is novel as it analysis not only unimodal systems but multimodal systems that combines three different datasets.

II. Literature Review

Authors have researched several methods for the detection of COVID-19 from images. For example, Houby [6] developed a deep learning model that extracts key features from X-ray images and then leverages a pre-trained model for the classification of COVID-19. A pre-trained model reflects an already existing model trained on thousands of images. Therefore, commonly preferred in image classification tasks—since it has learned images sufficiently from different domains. There is a high likelihood of improved performance when used in various classification tasks, including medical image classification tasks. Several pre-trained models have been used. According to the findings in Ref. [10], several pre-trained models were used for their proposed model. Some of the pre-trained models (transfer learning models) are ResNet-18, ResNet-50, and ResNet-101 [11]. The pre-trained models were also finetuned on the collected X-ray images, where the model trained on ResNET-101 proved to have outperformed other models. In addition, other research using deep learning models for X-ray COVID-19 detection has also been done as seen in Refs. [12, 13]. Haruna *et al.* [14] used the VGG-19 [15] pre-trained model for the classification of X-ray images with improved performance.

The CT-scan images have also been solely analyzed for the detection of the presence of the COVID-19 virus. For example, earlier research [16] used a similar approach proposed in Ref. [10]. However, they used pre-trained models built on the VGG-19 and then compared them to other pre-trained models such as Xception Net, and Convolutional Neural Network (CNN). Their analysis indicated satisfactory performance. In Ref. [17], a deep learning model was developed that improves on traditional deep learning methods. Their model incorporates two key innovations—the ability to reason based on the passed data, and the ability to learn. This means the model does not need input from humans in setting parameters. The model achieved an F1-Score of 0.9731 on CT-scan images [17–19].

Furthermore, evidence provided in Ref. [20] suggests that the adoption of a pre-trained model for CT-scan can further improve its classification output compared to models that do not use pre-trained models. On the other hand, researchers have also solely used cough datasets for the prediction of COVID-19. This is because patients suffering from the virus have a distinct way of coughing. This is because, their lung structure has been altered [21, 22]. A Support Vector Machine (SVM) was used in Ref. [23] to classify voice sounds. An ensemble model has also been proposed in Ref. [24]. The model consists of a CNN layer for feature extraction, and then another classification model. The authors went further to develop an application called "AI4COVID-19" where users can interface with their tool.

The papers reviewed above point to the fact that pre-trained models can potentially improve the output of COVID-19 classification. More recently, the use of multimodal deep learning models has been adopted to improve COVID-19 accuracy. For example, in Ref. [25], a multimodal system was developed that uses X-ray and CT scan images for classification. The study further experimented with different transfer learning architectures such as the MobileNetV2 [26], VGG-19, and ResNet-50 [25]. Similarly, two pre-trained models were used in Ref. [27], one for the CT scan and another for the X-ray. The outputs of these different deep learning layers are then fused to give an improved classification output. Many researchers have also combined datasets from different sources; for example, a combination of X-ray and cough datasets was used in Ref. [28] to improve classification. Consequently, two models were developed, and then the outputs of these models were fused. In addition, their cost function gives more weight or relevance to the model with the least error.

As shown in Ref. [29], the approach uses both X-ray and CT scan images for detection. From their experiment, it was established that VGG-19 gave the best result in terms of classification accuracy. It was also established that X-ray images are more accurate in the detection of the virus compared to CT-scan images.

While COVID-19 detection with multimodal architectures exists, as discussed, the research objective is to develop a multimodal architecture that takes advantage of cough, X-ray, and CT scans together for the classification of the COVID-19 virus. The second objective is to investigate the relevance of pre-trained models on multimodal and unimodal architectures using these three datasets.

The summary of our key contributions is as follows: To compare the prediction accuracy of COVID-19 using unimodal systems on individual datasets such as CT scan, cough sound, and X-ray with the prediction of COVID-19 using a combination (multimodal) of all the considered datasets. This is important as it provides an avenue for medical personnel to choose from an array of options that best aid objective COVID-19 diagnosis.

Our research objectives are to fill this gap by researching the relevance of a multimodal system using three datasets in the classification of COVID-19 with or without a pre-trained model. In addition, we also explored the contribution of pre-trained models on unimodal systems for cough, X-ray, and CT-scan datasets.

This research is important as it gives a sense of direction as to which of the datasets could potentially be used for the effective prediction of COVID-19.

III. Materials and Methods

Many researchers have extensively used unimodal pre-trained models to improve the classification of COVID-19 via X-ray images, CT-scan, and cough, as well as multimodal designs for different combinations such as X-ray and CT scan. Multimodal designs that combine cough, CT-scan, and X-rays remain largely unexplored in existing studies. In addition, we investigate further the impact of the VGG-19 pre-trained model on unimodal classifications such as for cough, CT-scan, and X-ray, and then the effect of the VGG-19 on a multimodal design that leverages three datasets—cough, X-ray and CT scan. Therefore, this research investigates the impact of pre-trained models (VGG-19, ResNET, and MobileNetV2) on a multimodal system that combines three datasets of cough, X-ray and CT scan.

This research is important as it provides an in-depth analysis of how the combination of three datasets can be used to improve COVID-19 classification. In addition, it provides a comparative study of how selected pre-trained models can influence COVID-19 classification from the perspective of unimodal and multimodal systems. Furthermore, this research also provides a platform to look at medical diagnosis from the multimodal perspective as it may potentially improve classification.

The motivation behind the combination of these datasets is based on the fact that COVID-19 patients usually exhibit these symptoms—lung abnormalities, and cough [30]. Lung abnormalities can be detected from X-ray images as well as CT scans. There are several advantages to the proposed model of using three datasets. One of them is the ability to harness complementary information—the sound from a cough could depict respiratory disease while images from X-rays and CT scans could reveal the structure of the lungs. In addition, X-rays and CT scans can also reveal the extent of damage to the lungs while the cough sounds might indicate early signs of COVID-19. In addition, there is the potential to increase sensitivity (correctly identifying persons with COVID-19) and specificity (correctly identifying persons without COVID-19) Lastly, data from one source might be unreliable for examination due to poor X-ray or CT-scan images [31].

This section discusses the architecture of the proposed model as well as the data preprocessing stages for the datasets used. Recall that we used three datasets: cough sound, X-ray, and CT scan. We therefore need to preprocess the data before passing it into the proposed multimodal architecture. The equations from this section have been derived from Pytorch framework [32].

A. Data Pre-Processing

The first dataset is the cough dataset found in Ref. [33]. This dataset has audio that lasts up to 9 seconds, however, on average, each audio contains approximately a two-second cough segment. Therefore, we needed to extract this segment. To do this, each cough file is loaded using the PyTorch torchaudio.load() function. This function then reads the audio waveform (*wf*) as seen in Eq. (1).

$$wf, sample\_rate = \text{torchaudio.load(audio\_path)} \quad (1)$$

In Eq. (1), wf is the cough waveform (*wf*). The next stage is resampling. If the sample rate of the audio is not at 16 kHz, we then resample (Eq. (2)) [34].

$$\text{wf} = \text{Resample(wf}, 16000) \quad (2)$$

The conversion of the waveform to a mono channel is next (Eq. (3)). The conversion is necessary because, the critical information (cough) can be found in one channel, another advantage of using one channel is for noise reduction [35].

$$\text{wf} = \text{wf}[0:1,:] \quad (3)$$

Next, the detection of the cough segment kicks in and it is then extracted. This is done by using the Short-Time Fourier Transform (STFT) [36, 37]. The STFT breaks the cough waveform into small time windows that would enable signal change detection [36]. STFT converts the waveform from a time domain into the frequency domain (F) for each frame $\text{F} = |\text{STFT}(wf)|$ [36, 37].

The next is to calculate the energy of each frame as seen in Eq. (4) [37]:

$$energy = \sum(F^2, axis = 0) \quad (4)$$

where F^2 squares the size or magnitude of the STFT for each frame. In the next stage, we normalized the energy (Eq. (5)) [37].

$$norm_{energy} = \frac{energy - \min(energy)}{\max(energy) - \min(energy)} \quad (5)$$

The normalized energy ($norm_{energy}$) is then analyzed based on a set threshold of 0.5. If an energy's frame exceeds 0.5, the algorithm then selects the first two seconds of the waveform, effectively collecting the cough portion into X_{audio} [37].

We used the X-ray dataset found in Refs. [38–44] while for the CT scan, we used the dataset in Refs. [17–19, 45]. The pre-processing for X-ray and CT-scan images is similar. Each image is opened using the PIL image.open() function [46], and read as coloured images. The images are further resized into 224 by 224—since some pre-trained models take images in this size. The preprocessed X-ray images are stored in X_{xray} while $X_{ct-scan}$ is for CT-scan.

B. Proposed Model

Fig. 1 depicts the proposed model which has five layers, excluding the pre-processing layer. The pre-processing layer ensures each dataset has a size of 224 by 224 by 3. For C1 (which houses the first convolution layer for cough). The preprocessed cough sound (X_audio) is being fed into the model via block C1. This block takes the processed cough sounds as input as seen in the equation below. Eqs. (6)–(8) encapsulate the processes in block C1. In Eq. (7) p is the dropout value, set to 0.5. in_channels are input neurons, while out_channels are output neurons. Kernel depicts the shape of the filter, while stride reflects

the kernel's movement. MaxPool2D helps in the shrinkage of the features as they move towards the classification layer. Padding of 1 means adding one pixel around the feature boundary. Dropout means randomly removes some neurons from the network [47].

The convolutional blocks used in Eqs. (6)–(27) follow the design patterns of CNNs [32, 48], which consist of convolutional layers, ReLU activations, dropout, Kernel (mask size) and max-pooling.

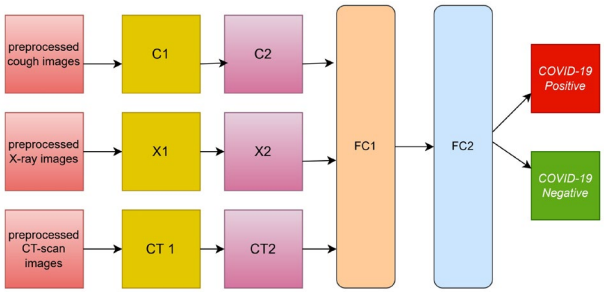


Fig. 1. Proposed multimodal architecture.

$$C_1 = \text{ReLU}\,(\text{Conv2D}(\text{in}_{\text{channel}} = 3, \text{out}_{\text{channels}}: 32, \text{mask size} = 3, \text{stride} = 1, \text{padding} = 1)) \quad (6)$$

$$C_1 = \text{Dropout}\,(C_1), p = 0.5) \quad (7)$$

$$C_1 = \text{MaxPool2D}\,(C_1, \text{mask size} = 2, \text{stride} = 2) \quad (8)$$

For block *C2,* we have Eqs. (9) and (10) where C_1 is the output from the previous layer while W_2 and b_2 are the weights and biases of the block.

$$C_2 = \text{ReLU}(\text{Conv2D}(C_1, \text{in}_{\text{channel}} = 32, \text{out}_{\text{channels}}: 64, \text{mask size} = 3, \text{stride} = 1, \text{padding} = 1)) \quad (9)$$

$$C_2 = \text{Dropout}\,(C_2, p = 0.5)$$

$$C_2 = \text{MaxPool2D}\,(C_2, \text{mask size} = 2, \text{stride} = 2) \quad (10)$$

For X-ray pre-processed images, we have Eqs. (11), (12)–(16) for blocks X1 and X2.

$$X_1 = \text{ReLU}(\text{Conv2D}(\text{in}_{\text{channel}} = 3, \text{out}_{\text{channels}}: 32, \text{mask size} = 3, \text{stride} = 1, \text{padding} = 1)) \quad (11)$$

$$X_1 = \text{Dropout}\,(X_1, p = 0.5) \quad (12)$$

$$X_1 = \text{MaxPool2D}\,(X_1, \text{mask size} = 2, \text{stride} = 2) \quad (13)$$

$$X_2 = \text{ReLU}(\text{Conv2D}(X_1, \text{in}_{\text{channel}} = 32, \text{out}_{\text{channels}}: 64, \text{mask size} = 3, \text{stride} = 1, \text{padding} = 1)) \quad (14)$$

$$X_2 = \text{Dropout}\,(X_2, p = 0.5) \quad (15)$$

$$X_2 = \text{MaxPool2D}\,(X_2, \text{mask size} = 2, \text{stide} = 2) \quad (16)$$

For CT-scan images, we have the following Eqs. (17)–(22).

$$CT_1 = \text{ReLU}\,(\text{in}_{\text{channel}} = 3, \text{out}_{\text{channels}}: 32, \text{mask size} = 3, \text{stride} = 1, \text{padding} = 1) \quad (17)$$

$$CT_1 = \text{Dropout}\,(CT_1, p = 0.5) \quad (18)$$

$$CT_1 = \text{MaxPool2D}\,(CT_1, \text{mask size} = 2, \text{stide} = 2) \quad (19)$$

$$CT_2 = \text{ReLU}\,(\text{Conv2D}(CT_1, \text{in}_{\text{channel}} = 32, \text{out}_{\text{channels}}: 64, \text{mask size} = 3, \text{stride} = 1, \text{padding} = 1)) \quad (20)$$

$$CT_2 = \text{Dropout}\,(CT_2, p = 0.5) \quad (21)$$

$$CT_2 = \text{MaxPool2D}\,(CT_2, \text{mask size} = 2, \text{stide} = 2) \quad (22)$$

For FC1, in Fig. 1, we have Eq. (23) that combines the outputs from C2, X2, and CT2. FC_1 is then passed onto FC_2 and then to the sigmoid function (Eq. (25)) for binary classification.

$$FC_1 = \text{ReLU}\left(\begin{bmatrix} C_2 \\ X_2 \\ CT_2 \end{bmatrix}, \text{in}_{\text{features}} = 602112, \text{out}_{\text{features}} = 164\right) \quad (23)$$

$$FC_2 = \sigma(FC_1, \text{in}_{\text{features}} = 164, \text{out}_{\text{features}} = 1) \quad (24)$$

$$Ouput = \sigma(FC_2) \quad (25)$$

$$if\ Ouput \geq T\ (T = 0.5): \quad Positive\ (Covid - 19\ present) \quad (26)$$

$$if\ Ouput < T\ (T = 0.5): \quad Negative\ (Covid - 19\ absent) \quad (27)$$

In summary, each architecture in Fig. 1 has its own convolutional layers (C1, X1, CT1, and C2, X2, CT2). These compartments extract features from each dataset. This information is then concatenated [49]. Meaning critical features from each modality are merged and then passed on to the classification layer. Obviously, each modality contributes uniquely to the classification of COVID-19 classification.

For the training process, we used Eq. (28) [50]. Each dataset has an input size of 224 by 224. We trained for 30 epochs with a learning rate of 0.0001 and a batch size of 8. The inputs to the proposed model go into the model at the same time. This means that a sample of positive COVID-19 cough, X-ray and CT scan goes into the model at once. The same applies to negative COVID-19 samples.

$$\theta_{t-1} = \theta_t - lr\ \frac{\delta L(M(X_{audio,}, X_{xray,}, X_{ct-scan,}), y)}{\delta \theta} \quad (28)$$

In Eq. (28), θ is the model parameter, while $(M(X_{audio,}\ Xxray, X_{ct-scan,})$ is the forward pass of the model. Lr is the learning rate. Lastly, $L(M(X_{audio,}\ Xxray, X_{ct-scan}), y)$ is the loss function.

Fig. 2 provides the structure of the dataset for training, testing and validation.

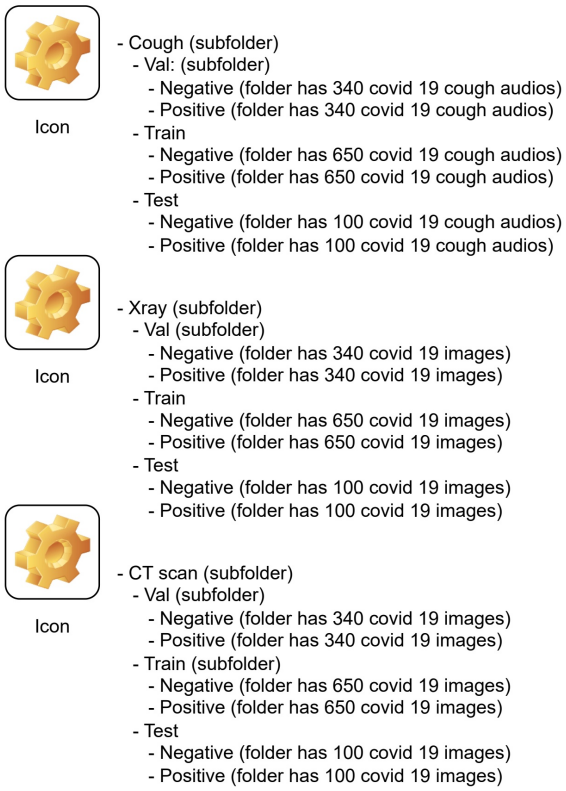


Fig. 2. Training, testing and validation configuration.

C. Validation

We perform validation on the model, and if the validation loss is lower than in the previous epoch, we save the current model state. For each batch, *i* of 8 data points (eight cough sounds, eight X-ray images, and eight CT-scan), do the following in Algorithm 1 [48, 51].

Algorithm 1: Model Loss Tracker

1. Initialize model's best loss $L_{best} = \infty$
2. Forward pass and compute loss
3. Compute accuracy
4. Then calculate the total loss $L^{(t)}$ on the validation dataset for each epoch.

$$L^{(t)} = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{loss_i}\right)$$

5. $if\ L^{(t)} < L_{best}, then\ L_{best} = L^{(t)}$

IV. Unimodal Systems

To test the robustness of the proposed model, we developed two versions of the model for each dataset, one without any pre-trained model (Fig. 3), and one using the VGG-19 pre-trained model (Fig. 4). In the unimodal system in Fig. 3, the T1 and T2 blocks cough replicate the C1 and C2 blocks in Fig. 1. Similarly, the T1 and T2 blocks for X-ray replicate the X1 and X2 blocks. The same applies to CT scans, where CT1 and CT2 correspond to T1 and T2. This also applies to FC1 and FC2.

In Fig. 4, the pre-trained model is the VGG-19. FC2 reflect the same architecture as seen in Fig. 1. While for FC1, in-features is 4096 and out-features is 164. The FC1 layer modifies the classification layer of the pre-trained model (VGG-19). The training, and testing validation of the unimodal systems follow the same process prescribed for the multimodal systems, earlier. Except that the combination of datasets is not implemented. Each of the pre-trained models used in this study was fine-tuned on the target dataset.

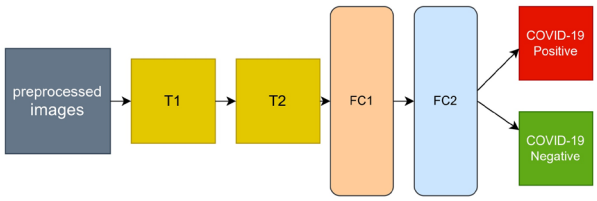


Fig. 3. Unimodal architecture used for Cough, X-ray, and CT-scan for COVID-19 classification.

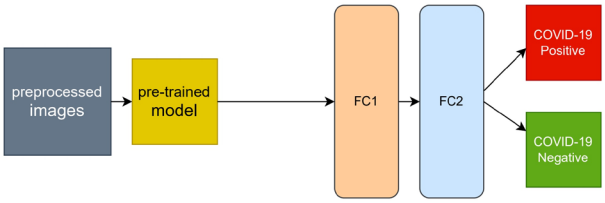


Fig. 4. Unimodal architecture with pre-trained model (VGG-19) used for cough, X-ray, and CT-scan for COVID-19 classification.

V. Multimodal Pre-Trained Systems

Furthermore, we developed a multimodal pre-trained model (Fig. 5). Its FC1 layer has 31,360 input features and 64 out features (VGG-19). FC2 has 64 input features and 1 output feature (VGG-19). We experimented with three pre-trained models: VGG-19, ResNet-18 and MobileNetV2, as they are widely used in the literature for image classification tasks.

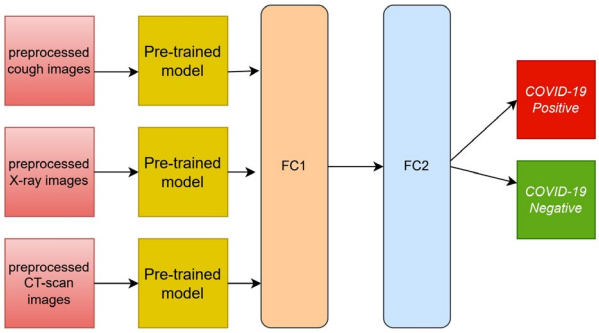


Fig. 5. Multimodal architecture with pre-trained model (VGG19) used for cough, X-ray, and CT-scan for COVID-19 classification.

VI. Evaluation

To evaluate the proposed multimodal system, we passed test dataset samples in batches of 8. Each sample includes a cough signal (C), an X-ray image (X), and a CT scan (CT). These inputs are then fed into the model to generate a prediction: $y_i = M(C_I, X_I, CT_I)$ where M represents the trained model. We also performed this evaluation on the multimodal system with pre-trained models, as well as on unimodal systems with and without pre-trained models. The following evaluation metrics are used (Eqs. (29)–(33)), where TP = true positive, TN = true negative, FP = false positive, and FN = false negative.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100 \quad (29)$$

$$Sensitivity = \frac{TP}{TP+FN} \quad (30)$$

$$Specificity = \frac{TN}{TN+FP} \quad (31)$$

$$F1-score = 2 \times \frac{Precision \times Recall}{TN+FP} \quad (32)$$

$$Precision = \frac{TP}{TN+FP} \quad (33)$$

VII. Result

The results highlight several key insights into the use of deep learning models. First, for the unimodal systems for cough, X-ray, and CT scan, the result from Table I shows the accuracy score as well as the F1-Score of the unimodal system for cough. Looking at the graphs of the unimodal and pre-trained unimodal models in Figs. 6 and 7, it is clear that the model without a pre-trained model (Fig. 6) shows a flat validation loss throughout. This suggests it may not be learning during training, which could explain the poor results (Table I). In Table II, there is improved performance in accuracy. This shows the importance of transfer learning on the cough dataset in delivering a better outcome. The architectural complexity of VGG-19 might have contributed to this improvement.

The same outcome is also observed for the unimodal deep learning system for X-ray (Tables III and IV, Figs. 8 and 9); however, it shows an improved F1-Score in addition to the accuracy metric. The pre-trained VGG-19 model outperformed the traditional CNN-based deep learning model (Tables III and IV). It can also be observed that overfitting is minimized (Fig. 9), as the training and validation loss curves are relatively close.

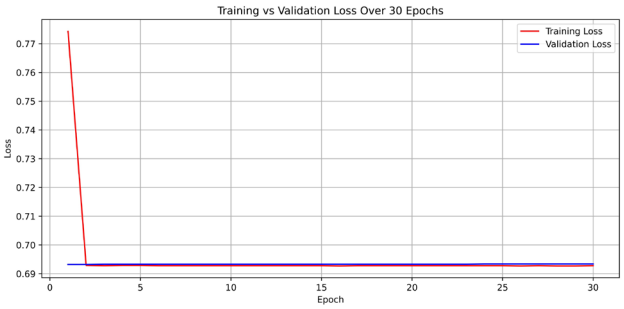


Fig. 6. Unimodal COVID-19 cough classification—training loss vs validation loss.

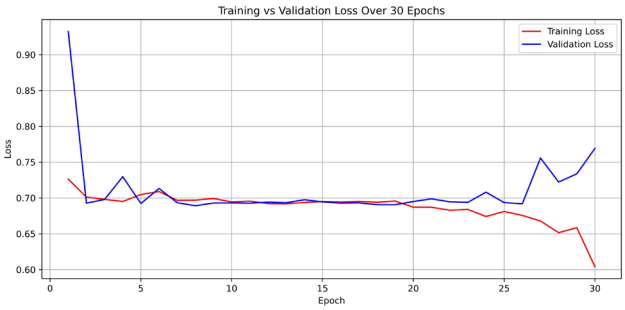


Fig. 7. Unimodal COVID-19 cough classification with VGG-19—training loss vs validation loss.

TABLE I. Evaluation for Unimodal–Cough

Evaluation metrics	Values
Accuracy	50%
Sensitivity	1
Specificity	0
F1-Score	0.6667
Confusion matrix	[[0 100] [0 100]]
True Positives (TP)	100
True Negatives (TN)	0
False Positives (FP)	100
False Negatives (FN)	0

TABLE II. Evaluation for Unimodal–Cough VGG-19

Evaluation metrics	Values
Accuracy	55%
Sensitivity	0.68
Specificity	0.42
F1-Score	0.6018
Confusion matrix	[[42 58] [32 68]]
TP	68
TN	42
FP	58
FN	32

TABLE III. Evaluation for Unimodal–X-Ray

Evaluation metrics	Values
Accuracy	98.00%
Sensitivity	1
Specificity	0.960
F1-Score	0.9804
Confusion matrix	[[96 4] [0 100]]
TP	100
TN	96
FP	4
FN	0

TABLE IV. Evaluation for Unimodal–X-Ray VGG-19

Evaluation metrics	Values
Accuracy	99.00%
Sensitivity	1
Specificity	0.98
F1-Score	0.9901
Confusion matrix	[[98 2] [0 100]]
TP	100
TN	98
FP	2
FN	0

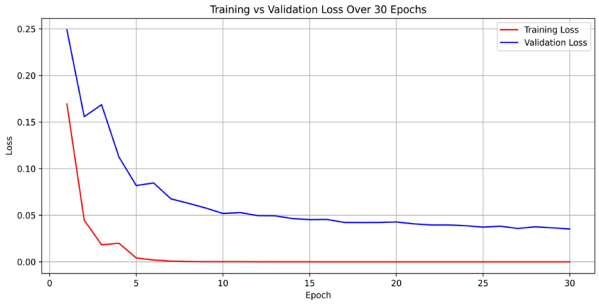


Fig. 8. Unimodal COVID-19 X-ray classification—training loss vs validation loss.

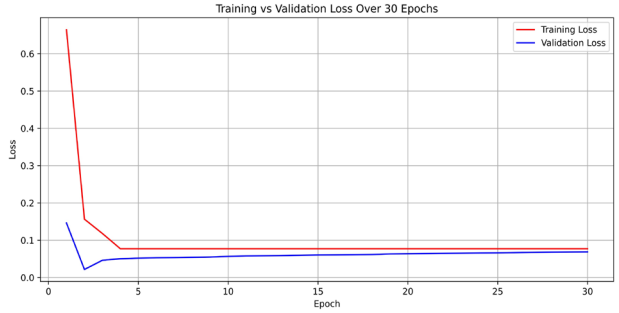


Fig. 9. Unimodal COVID-19 X-ray classification with VGG-19—training loss vs validation loss.

However, this is not the case for the CT scan. The traditional unimodal deep learning model outperformed the pre-trained unimodal model for CT scans (Tables V and VI) in almost all the metrics. One possible explanation is that the images learned from the pre-trained model may not generalize well to CT scan data. Both the pre-trained and traditional models did attempt to address the overfitting issue during training (Figs. 10 and 11).

TABLE V. EVALUATION FOR UNIMODAL–CT-SCAN

Evaluation metrics	Values
Accuracy	91.00%
Sensitivity	0.91
Specificity	0.91
F1-Score	0.91
Confusion matrix	[[91 9] [9 91]]
TP	91
TN	91
FP	9
FN	9

TABLE VI. EVALUATION FOR UNIMODAL–CT-SCAN VGG-19

Evaluation metrics	Values
Accuracy	88.50%
Sensitivity	0.80
Specificity	0.97
F1-Score	0.8743
Confusion matrix	[[97 3] [20 80]]
TP	80
TN	97
FP	3
FN	20

Fig. 10. Unimodal COVID-19 CT-scan classification—training loss vs validation loss.

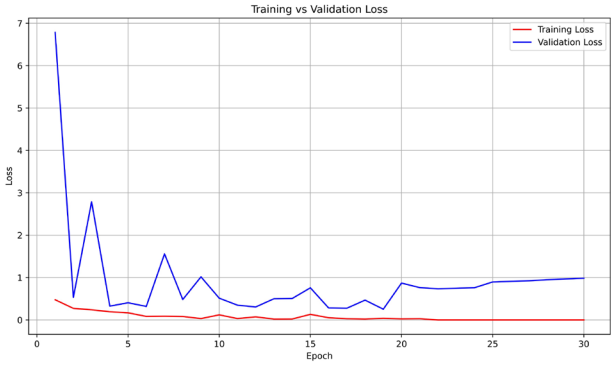


Fig. 11. Unimodal COVID-19 CT-scan classification with VGG-19—training loss vs validation loss.

Moving on to multimodal systems (Tables VII–X), it is observed that the model leveraging a pre-trained model and the one without both have the same accuracy of 98% (Tables VII vs. VIII). However, using the F1-Score, the non-pre-trained model outperformed the pre-trained model. This could be explained by the fact that the multimodal system has learned from three different datasets and aggregated complementary information from them. As a result, it is possible to learn unique attributes from these sources, enabling a more effective COVID-19 classification model.

We also extended the experiment to two additional pre-trained models—ResNet-18 and MobileNetV2—as shown in Tables IX and X; however, neither could outperform the multimodal system developed without a pre-trained model.

Using the F1-Score, it is also evident that the non-pre-trained multimodal model performs better than the unimodal CT scan model proposed in Refs. [17–19], which reported an F1-Score of 0.9731. The multimodal system we proposed is enriched by learning from three diverse datasets. In addition, analyzing the training graphs (Figs. 12 and 13) shows that the non-pre-trained multimodal system (Fig. 12) exhibits less overfitting compared to the VGG-19-based pre-trained multimodal system (Fig. 13).

TABLE VII. EVALUATION FOR PROPOSED MODEL–MULTIMODAL

Evaluation metrics	Values
Accuracy	98.00%
Sensitivity	1
Specificity	0.9600
F1-Score	0.9804
Confusion matrix	[[96 4] [0 100]]
TP	100
TN	96
FP	4
FN	0

In Table I, it is evident that the unimodal system without the pre-trained model correctly identified COVID-19 positive cases (as sensitivity is 1); however, it failed to identify non-COVID-19 cases (specificity is 0). Meanwhile, in Table II, the unimodal system equipped with VGG-19 strikes a balance—sensitivity is 0.68 while specificity is 0.42. Due to the introduction of the

pre-trained model, Table II shows improved results over Table I. It is important to note that in medical diagnosis, high sensitivity ensures that false negatives are significantly reduced. The implication of high sensitivity is a reduced likelihood of predicting that a person does not have COVID-19 when they actually do.

TABLE VIII. EVALUATION FOR PROPOSED MODE–MULTIMODAL VGG-19

Evaluation metrics	Values
Accuracy	98.00 %
Sensitivity	0.98
Specificity	0.98
F1-Score	0.980
Confusion matrix	[[98 2] [2 98]]
TP	98
TN	98
FP	2
FN	2

TABLE IX. EVALUATION FOR PROPOSED MODEL–MULTIMODAL RESNET-18

Evaluation metrics	Values
Accuracy	50.00 %
Sensitivity	1
Specificity	0
F1-Score	0.6667
Confusion matrix	[[0 100] [0 100]]
TP	100
TN	0
FP	100
FN	0

TABLE X. EVALUATION FOR PROPOSED MODEL–MULTIMODAL MOBILENETV2

Evaluation metrics	Values
Accuracy	96%
Sensitivity	0.96
Specificity	0.96
F1-Score	0.96
Confusion matrix	[[96 4] [4 96]]
TP	96
TN	96
FP	4
FN	4

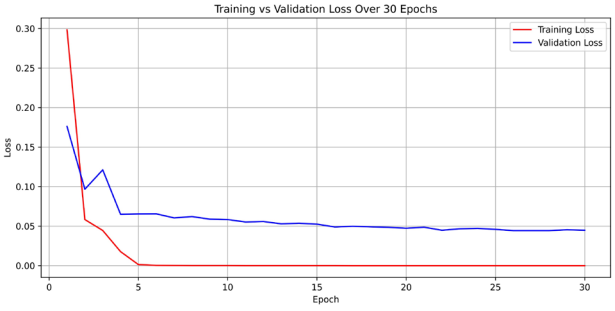


Fig. 12. Multimodal COVID-19 classification - training loss vs validation loss.

The same trend is also observed for X-ray (Tables III and IV). It is observed that the VGG-19 intervention improved the specificity score. However, for CT scan (Tables V and VI), the reverse is the case—the model without VGG-19 performed better in terms of sensitivity, accuracy and F1-Score. This shows that the VGG-19 pre-trained model did not complement the CT scan images.

Furthermore, for the multimodal system (Table VII) without VGG-19, sensitivity is 1, while its specificity is 0.96. However, for the multimodal system with VGG-19 (Table VIII), sensitivity is 0.98, and its specificity is 0.98. The multimodal system without VGG-19 shows improved sensitivity—meaning it has zero false negatives.

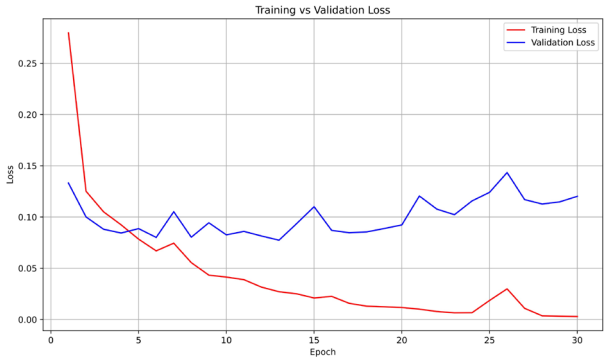
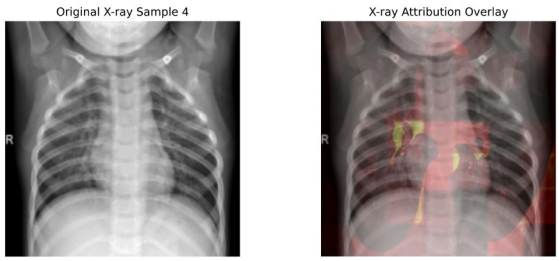
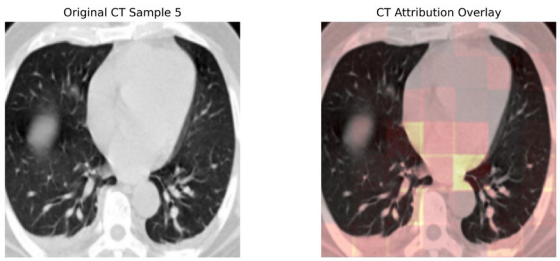


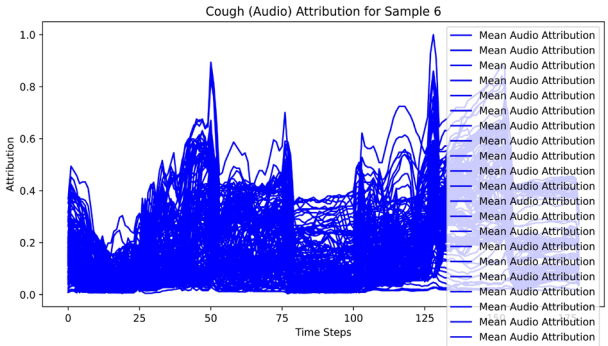
Fig. 13. Multimodal COVID-19 classification with VGG-19 - training loss vs validation loss.



(a)



(b)



(c)

Fig. 14. How datasets contribute to COVID-19 prediction. (a) X-ray. (b) CT-scan (c). Cough.

Fig. 14 shows how each dataset contributes to the prediction of COVID-19. Figs. 14(a) and 14(b) show bright areas. These bright areas are critical structures that are used to make decisions. In addition, Fig. 14(c) shows spikes with higher amplitudes. This indicates that specific spikes at given times also contributed to the COVID-19 prediction. We used the framework from Refs. [52, 53] to generate the diagram shown. The diagram illustrates how the three datasets contributed to COVID-19 prediction. A similar approach was also adopted in Refs. [54−57].

VIII. Discussions

The results highlight several key insights into the use of deep learning models. First, for the unimodal systems for cough, X-ray, and CT-scan: the result from Table I shows the accuracy score as well as the F1-Score of the unimodal system for cough. In Table II, there is an improved performance in accuracy. This shows the importance of transfer learning on the cough dataset to deliver an improved outcome. A similar outcome is recorded for unimodal deep learning systems using X-rays. The pre-trained model for the VGG-19 outperformed the traditional CNN deep learning model (Tables III and IV) in both accuracy and F1-Score. However, this is not the case for the CT scan. The unimodal deep learning model mostly outperformed the pre-trained unimodal for CT-scans (Tables V and VI). An explanation for this is that the images used on the pre-trained models did not complement the CT-scan images.

VGG-19 pre-trained models are trained on thousands of images, where edges and structures may be similar to those found in X-ray images and the visual representation of cough. Since there is an improvement in the output of pre-trained unimodal models for cough and X-ray, it suggests that the structures and textures learned by VGG-19 complement these unimodal datasets. However, for CT scans, the edges and textures of the images used in the pre-training process differ significantly from those in CT-scan images. For this reason, the CT-scan model backed with a pre-trained VGG-19 did not perform well.

Moving on to multimodal systems, it is observed that models that leveraged a pre-trained model and those without a pre-trained model are close in terms of F1-Score. Using the F1-Score, the model without the pre-trained model outperformed (0.9804 F1-Score) the pre-trained model (0.98 F1-Score). An explanation for this is that since the multimodal system has learned from three different datasets, it aggregated complementary information from them. As a result, there is a high possibility that it learned the unique attributes from these datasets to deliver an improved COVID-19 model. On the other hand, for the pre-trained multimodal system, having already learned the unique features from these three data sources was sufficient. Therefore, the pre-trained model was not necessary. This claim could also be observed in the other experiments with ResNET-18 and MobileNetV2. In Tables IX and X, these pre-trained models could not outperform the multimodal system developed without a pre-trained model.

The multimodal system is enriched by learning from three diverse datasets. From Fig. 13, we can observe that the multimodal system with a pre-trained model over-fits. This also shows that the VGG-19 pre-trained model was learning almost entirely from its own training dataset and not adapting to the new dataset. This contrasts with the non-pre-trained multimodal model (Fig. 12). These findings show that for resource-constrained economies that may not have the capacity to acquire expensive CT scan equipment, a unimodal system (equipped with VGG-19) that takes input from an X-ray for the prediction of COVID-19 would be effective. On the other hand, when funds are not limited, investment in a multimodal system is recommended, as it can learn features from various datasets to ensure an objective diagnosis. While the training images for the models investigated were limited to 1300 per dataset, it would be interesting to see the performance when the training dataset is increased to about 5000.

IX. Conclusion

This study developed a multimodal deep learning system for classifying COVID-19 using three datasets—cough, X-ray, and CT scan. Using pre-trained models such as ResNet-18, VGG-19, and MobileNetV2, the results show that a multimodal system combining these datasets can deliver improved performance even without pre-trained models. This is possible because the multimodal system has learned sufficiently from different heterogeneous datasets, making it robust enough to perform well without a pre-trained model. This is also reflected in the sensitivity metric: the non-pre-trained multimodal system scored 1, while the pre-trained model scored 0.98. High sensitivity indicates fewer false negatives in COVID-19 prediction. However, this is not the case for the unimodal models developed for cough and X-ray—their pre-trained versions strike a balance between sensitivity and specificity, reducing both false negatives and false positives.

An explanation for this is that the features learned from VGG-19 complement the training process. On the flip side, this was not the case for the unimodal deep learning model developed for the CT scan (based on F1-Score, sensitivity, and accuracy). It was discovered that the pre-trained model did not outperform the non-pre-trained version. A possible explanation is that the features learned by the pre-trained model do not align with CT scan images. One possible solution is to investigate, from the pool of available pre-trained models, which model best aligns with or improves COVID-19 classification from CT scans. Another option is to apply image preprocessing to enhance the appearance of CT scan images. This will be explored in our future work.

The importance of this research is that it provides a platform for researchers and medical experts to identify which dataset combinations could be considered, given the level of resources at their disposal. For example, a medical expert might opt for a system with high sensitivity or high specificity and then decide which model to adopt based on available resources.

In summary, for unimodal systems, an X-ray with a pre-trained VGG-19 model could be reliable it would give higher accuracy and F1-Score compared to the scores of the CT scan or cough models. However, for multimodal systems, a non-pre-trained model yields a better F1-Score. Deploying the models requires either a cloud platform or a dedicated computer system hosted locally to handle service requests. This means that medical facilities with limited resources might find it difficult to acquire the necessary hardware or cloud infrastructure to run a multimodal system. In such cases, a unimodal system could be considered.

Also, in terms of noise, an improved data preprocessing method needs to be integrated for CT scans or X-rays. A preprocessing layer that incorporates a filter for noise removal could be included. For cough noise, a preprocessing layer that extracts the segmented cough, as illustrated earlier, is essential. Regarding patient variability which could potentially affect the model's performance when deployed in different environments large and diverse datasets are required to sufficiently capture this variability.

While this paper focuses on three datasets X-ray, CT scan, and cough sound the integration of medical history and blood test results could further improve prediction accuracy. In this case, the developed multimodal model will be adjusted to handle text-based data from test results and patient history. This implies incorporating Natural Language Processing (NLP) into the model, which we plan to explore in future research.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

KO developed the models algorithm together with EE; KO did the documentation together with EE; MC did the final editing and correction; all authors had approved the final version.

FUNDING

This work is partly funded by the Centre for Artificial Intelligence and Multidisciplinary Innovation Studies, College of Accounting Science, University of South Africa under the office of the Executive Deputy Dean.

REFERENCES

[1] B. Hu, H. Guo, P. Zhou *et al.*, "Characteristics of SARS-CoV-2 and COVID-19," *Nature Reviews Microbiology*, vol. 19, no. 3, pp. 141–154, 2021.

[2] WHO. (Dec 2024). WHO COVID-19 Dashboard: Deaths. *World Health Organization*. [Online]. Available: https://data.who.int/dashboards/covid19/deaths

[3] Y. J. Chen, W. H. Jian, Z. Y. Liang *et al.*, "Earlier diagnosis improves COVID-19 prognosis: A nationwide retrospective cohort analysis," *Ann. Transl. Med.*, vol. 9, no. 11, 941, 2021.

[4] L. J. Layfield, S. Camp, K. Bowers *et al.*, "SARS-CoV-2 detection by reverse transcriptase polymerase chain reaction testing: Analysis of false positive results and recommendations for quality control measures," *Pathology - Research and Practice*, vol. 225, 153579, 2021.

[5] M. Teymouri, S. Mollazadeh, H. Mortazavi *et al.*, "Recent advances and challenges of RT-PCR tests for the diagnosis of COVID19," *Pathology - Research and Practice*, vol. 221, 153443, 2021.

[6] E. M. F. E. Houby, "COVID-19 detection from chest X-ray images using transfer learning," *Scientific Reports*, vol. 14, 11639, 2024.

[7] R. Fusco, R. Grassi, V. Granata *et al.*, "Artificial intelligence and COVID-19 using chest CT scan and chest X-ray images: Machine learning and deep learning approaches for diagnosis and treatment," *J. Pers. Med.*, vol. 11, no. 10, 993, 2021.

[8] M. Melek, "Diagnosis of COVID-19 and non-COVID-19 patients by classifying only a single cough sound," *Neural Comput. Appl.*, vol. 33, no. 24, pp. 17621–17632, 2021.

[9] K. Nguyen-Trong and K. Nguyen-Hoang, "Multi-modal approach for COVID-19 detection using coughs and self-reported symptoms," *Journal of Intelligent & Fuzzy Systems*, vol. 44, no. 3, pp. 3501–3513, 2023.

[10] M. Constantinou, T. Exarchos, A. G. Vrahatis *et al.*, "COVID-19 classification on chest X-ray images using deep learning methods," *Int. J. Environ. Res. Public Health*, vol. 20, no. 3, 2035, 2023.

[11] K. He, X. Zhang, S. Ren *et al.*, "Deep residual learning for image recognition," in *Proc. 2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, 2016.

[12] Z. Li, Q. Xing, J. Zhao *et al.,* "COVID19-ResCapsNet: A novel residual capsule network for COVID-19 detection from chest X-ray scans images," *IEEE Access*, vol. 11, pp. 52923–52937, 2023.

[13] M. Nahiduzzaman, M. O. F. Goni, M. S. Anower *et al.*, "A novel method for multivariant pneumonia classification based on hybrid CNN-PCA based feature extraction using extreme learning machine with CXR images," *IEEE Access*, vol. 9, pp. 147512–147526, 2021.

[14] U. Haruna, R. Ali, and M. Man, "A new modification CNN using VGG19 and ResNet50V2 for classification of COVID-19 from X-ray radiograph images," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 31, no. 1, pp. 369–377, 2023.

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

[16] K. K. Mohbey, S. Sharma, S. Kumar *et al.*, "COVID-19 identification and analysis using CT scan images: Deep transfer learning-based approach," *Blockchain Applications for Healthcare Informatics*, pp. 447–470, 2022.

[17] E. Soares, P. Angelov, S. Biaso *et al.*, "SARS-CoV-2 CT-scan dataset: A large dataset of real patients CT scans for SARS-CoV-2 identification," medRxiv Preprint, 2020.

[18] P. Angelov and E. A. Soares, "Explainable-by-design approach for COVID-19 classification via CT-scan," medRxiv Preprint, 2020. doi: https://doi.org/10.1101/2020.04.24.20078584

[19] P. Angelov and E. Soares, "Towards Explainable Deep Neural Networks (xDNN)," arXiv Preprint, arXiv: 1912.025232019, 2019.

[20] W. Zhao, W. Jiang, and X. Qiu, "Deep learning for COVID-19 detection based on CT images," *Scientific Reports*, vol. 11, no. 1, 14353, 2021.

[21] R. Han, L. Huang, H. Jiang *et al.*, "Early clinical and CT manifestations of coronavirus disease 2019 (COVID-19) pneumonia," *American Journal of Roentgenology*, vol. 215, no. 2, pp. 338–343, 2020.

[22] WHO. (March 2020). Clinical management of severe acute respiratory infection (SARI) when COVID-19 disease is suspected: interim guidance. *World Health Organization*. [Online]. Available: https://iris.who.int/handle/10665/331446

[23] J. Han, C. Brown, J. Chauhan *et al.*, "Exploring automatic COVID-19 diagnosis via voice and symptoms from crowdsourced data," in *Proc. IEEE International Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 8328–8332.

[24] A. Imran, I. Posokhova, H. N. Qureshi *et al.*, "AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app," *Informatics in Medicine Unlocked*, vol. 20, no. 10, 100378, 2020.

[25] N. Nasir, A. Kansal, F. Barneih *et al.*, "Multi-modal image classification of COVID-19 cases using computed tomography and X-rays scans," *Intelligent Systems with Applications*, vol. 15, 200160, 2023.

[26] M. Sandler, A. Howard, M. Zhu *et al.*, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, 2018.

[27] N. Hilmizen, A. Bustamam, and D. Sarwinda, "The multimodal deep learning for diagnosing COVID-19 pneumonia from chest CT-scan and X-ray images," in *Proc. 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, Yogyakarta, 2020.
[28] S. Kumar, M. K. Chaube, S. H. Alsamhi *et al.*, "A novel multimodal fusion framework for early diagnosis and accurate classification of COVID-19 patients using X-ray images and speech signal processing techniques," *Comput. Methods Programs Biomed*, vol. 226, 107109, 2022.
[29] S. E. Mukhi, R. T. Varshini, and S. E. F. Sherley, "Diagnosis of COVID-19 from multimodal imaging data using optimized deep learning techniques," *SN Comput. Sci.*, vol. 4, 212, 2023.
[30] E. Jangam, A. A. D. Barreto, and C. S. R. Annavarapu, "Automatic detection of COVID-19 from chest automatic detection of COVID-19 from chest CT scan and chest X-Rays images using deep learning, transfer learning and stacking," *Appl. Intell.*, vol. 52, pp. 2243–2259, 2022.
[31] C. Mahanty, S. G. K. Patro, S. Rathor *et al.*, "A comprehensive review on COVID-19 detection based on cough sounds, symptoms, CXR, and CT images," *IEEE Access*, vol. 12, pp. 75412–75425, 2024.
[32] PyTorch Contributors. PyTorch Documentation. *PyTorch*. [Online]. Available: https://pytorch.org/docs/stable/index.html
[33] L. Orlandic, T. Teijeiro, and D. Atienza, "The COUGHVID crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms," *Scientific Data*, vol. 8, 156, 2021.
[34] P. Team. torchaudio.transforms.Resample. [Online]. Available: https://pytorch.org
[35] E. D. Haritaoglu, N. Rasmussen, D. C. H. Tan *et al.*, "Using deep learning with large aggregated datasets for COVID 19 classification from cough," arXiv Preprint, arXiv:2201.01669, 2022.
[36] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.
[37] J. Brady, V. Chandiramani, M. Chatwin *et al.*, "Can audio recordings be used to detect leaks and coughs during mechanical insufflation-exsufflation (MI E) treatment?" arXiv Preprint, arXiv:2504.05005, 2025.
[38] P. Prashant. (2020). Chest X-ray Images (COVID-19, Pneumonia). [Online]. Available: https://www.kaggle.com/datasets/prashant268/chest-xray-covid19-pneumonia
[39] J. P. Cohen, P. Morrison, L. Dao *et al.*, (2020). COVID-19 image data collection: Prospective predictions are the future. [Online]. Available: https://github.com/ieee8023/covid-chestxray-dataset
[40] G. Chung. (2020). Figure 1 COVID Chest X-ray Dataset. [Online]. Available: https://github.com/agchung/Figure1-COVID-chestxray-dataset
[41] Jtiptj. (2020). Chest X-ray Pneumonia COVID-19 Tuberculosis Dataset. [Online]. Available: https://www.kaggle.com/datasets/jtiptj/chest-xray-pneumoniacovid19tuberculosis
[42] T. Rahman, A. Khandakar, M. A. Kadir *et al.*, "Reliable tuberculosis detection using chest X-ray with deep learning, segmentation and visualization," *IEEE Access*, vol. 8, pp. 191586–191601, 2020.
[43] J. P. Cohen, P. Morrison, L. Dao *et al.*, "COVID 19 image data collection: Prospective predictions are the future," arXiv Preprint, arXiv:2006.11988, 2020.
[44] D. S. Kermany, M. Goldbaum, W. Cai *et al*., "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131.e9, 2018.
[45] P, Angelov and E. Soares, "Towards explainable deep neural networks (xDNN)," *Neural Networks*, vol. 130, pp. 185–194, 2020.
[46] A. Clark and Contributors. Pillow (PIL Fork) Documentation. *Pillow Docs*, [Online]. Available: https://pillow.readthedocs.io/en/stable/
[47] S. Albawi, T. A. Mohammed, and S. Al-Zawi, "Understanding of a convolutional neural network," in *Proc. International Conf. on Engineering and Technology (ICET),* Antalya, 2017.
[48] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, Cambridge: MIT Press, 2016.
[49] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," arXiv Preprint, arXiv:1705.09406, 2017.
[50] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.
[51] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 12, 2012.
[52] N. Kokhlikyan, V. Miglani, M. Martin *et al*., "Captum: A unified model interpretability library for PyTorch," arXiv Preprint, arXiv:2009.07896, 2020.
[53] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. International Conf. on Machine Learning (ICML)*, 2017, pp. 3319–3328.
[54] S. Solayman, S. A. Aumi, C. S. Mery *et al*., "Automatic COVID-19 prediction using explainable machine learning techniques," *Int. J. Cogn. Comput. Eng.*, vol. 4, pp. 36–46, 2023.
[55] S. Al. Ebiaredoh-Mienye, T. G. Swart, E. Esenogho *et al*., "A machine learning method with filter-based feature selection for improved detection of chronic kidney disease" *Bioengineering,* vol. 9, no. 8, 350, 2022. https://doi.org/10.3390/bioengineering9080350
[56] S. A. Ebiaredoh-Mienye, "Improved machine learning methods for classification of imbalanced data," M.S. thesis, Univ. of Johannesburg, Johannesburg, South Africa, 2021.
[57] G. Obaido, B. Ogbuokiri, T. G. Swart *et al*., "An interpretable machine learning approach for hepatitis B diagnosis" *Applied Sciences*, vol. 12, no. 21, 11127, 2022. doi: https://doi.org/10.3390/app122111127